International Journals of Advanced Research in Computer Science and Software Engineering
ISSN: 2277-128X (Volume-9, Issue-9)

Research Article | **September 2019**

Hybrid Learning Approach Based Aspect Category Detection for Sentiment Summarization with Co-Occurrence Data

[1]CH.Sravani, [2]Y. Ramu
[1]M.Tech Student, [2]Associate Professsor
Dept of Computer Science and Engineering
Shri Vishnnu Engineering College for Women, Bhimavaram, AndhraPradesh, India.

Abstract- User-generated reviews are precious decision-making resources. Identifying the feature categories mentioned in a specified review phrase (e.g. "food" and "service" in restaurant reviews) is a significant task for analyzing sentiment and mining opinion. Most prior researchers hold hand-crafted characteristics and a classification algorithm to achieve the assignment given a predefined aspect category set. The key step to achieve better efficiency is feature engineering that consumes a great deal of human effort and can be volatile when the product domain changes. A hybrid learning method is suggested in this project to automatically learn helpful characteristics for the identification of aspect categories. Specifically, on a big collection of reviews with noisy labels, a Hybrid Aspect Analysis Algorithm is first suggested to achieve ongoing word depictions. We subsequently suggest generating deeper and hybrid characteristics through the stacked neural networks on the word vectors. Finally, a logistic regression classifier is trained to predict the aspect category with hybrid characteristics. The tests are conducted on a SemEval-2014 benchmark dataset. In this paper we achieves the state of the art results with the F1 score of 90.10% on the dataset. Overall, our approach to representation learning outperforms traditional hand-crafted characteristics and embedding algorithms with current words.

Keywords: Electronic Reviews, Sentiment Analysis, Natural Language Processing, Opinion Mining, Machine Learning.

1. Introduction

User-generated reviews play a key role in the decision-making process of each individual. Opinion mining and sentiment analysis for internet reviews have become a subject of trendy studies in scholarly and industrial areas since the beginning of 2000 **[1]**.Aspect category detection is one of the mining duties of view that seeks to define the categories of aspect mentioned in a review phrase. Usually a set of aspect classifications is predefined making the job a multi-label classification issue. For instance, the aspect category set for restaurant reviews is described in SemEval-2014,{ "service," "food," "cost," "ambience," "anecdote / miscellaneous"}. "Service is top notch" should be identified as the aspect category in the sentence "Service is top notch."Opinion that is of restricted use without understanding the target **[2]**. Identifying the category of aspect helps to achieve a target-dependent feeling and adds to a summary of views specific to the aspect.

Previous research has suggested several models for this assignment, and classification of SVM is one of the most popular**[3]**. The importance of lexical data in aspect category identification has been demonstrated by these current techniques. However, the characteristics based on unigrams or n-gram generally use one-hot depictions and fail to capture semantine relationships between distinct words. Unless it appears in the test data, words that appear in the training data cannot provide any information. Associations between different words cannot be evaluated quantitatively through one-hot vectors. we suggest a representation learning strategy for aspect category identification to overcome the weaknesses of the current research.

First, wesuggest an embedding algorithm semi-supervised word. It depicts semantic relationships between words, word-aspect relationships, and relationships between words and elements of feeling. We average all the word vectors in a phrase after acquiring the word vectors as its constant representation**[4]**.Unlike current works that directly learn supervised classifiers based on phrase vectors**[5]**, we suggest creating deeper and hybrid characteristics that assist increase efficiency. For learning shared characteristics and aspect-specific characteristics respectively, two distinct types of neural networks are used. By combining them, we get the hybrid characteristics. The hybrid-trained logistic regression classifier achieves the state-of-the-art results on the SemEval-2014 benchmark dataset. The output, as well as a few powerful baselines, is greater than that of the best participating team.

The study's major contributions are summarized as follows:-

1. For aspect category identification, we suggest a representation learning strategy that achieves state-of - the-art efficiency on a benchmark dataset.
2. We suggest a semi-supervised word embedding algorithm capturing semantic relationships in a unified structure between phrases, word-aspect relationships, and sentiment-aspect relationships.
3. By using two distinct types of neural networks, we produce deeper and hybrid characteristics. It demonstrates better efficiency than either the characteristics shared or the characteristics specific to the aspect.

2. Related Work

Analysis of feelings based on aspect is a job of fine-grained opinion mining. The opinion goal can be broken down into entity and its elements in product reviews. Analysis of feelings based on elements is aimed at finding the elements and the respective feeling for them**[6]**. Theme modeling has become the mainstreaming approach to dealing with the issue in latest years. These techniques extract elements and categorize them into multiple subjects at the same time **[7].** The multi-grain theme model has been suggested. The model utilizes the worldwide subject to capture aspect-independent words and to capture aspect-specific phrases using local subjects **[8].**Expanded into distinct subjects the multi-grain topic model and divided aspect words and associated sentiment words. Detection of aspect categories is a unique case of assessment of aspect-based feelings. An aspect category set is provided in advance instead of extracting elements and the objective is to classify each review phrase into one or more aspect categories.**[9]** suggested a discriminatory model to predict the aspect of the product. To encode the word association, they used two types of parameters. One of them learns which words are linked to each aspect. The other learns which words are linked to each rating of stars. Learning the depictions of vector spaces for natural language texts has been successful in capturing fine-grained semantic and syntactic connections. **[10]** suggested a neural network language model that concurrently learned a distributed representation of each word together with the word sequence probability function.The term embedding has subsequently become a hot topic of studies to represent semantics in a distributed manner**[11]**. In addition to the unsupervised word embedding algorithms, learning task-specific word embedding has demonstrated successful efficiency in many assignments.**[12]**

3. Proposed Work

First, we suggest an embedding algorithm semi-supervised word. Fig 1 depicts semantic relationships between words, word-aspect relationships, and relationships between words and elements of feeling. we average all the word vectors in a phrase after acquiring the word vectors as their ongoing representation. Aspect category detection is a significant assessment task of SemEval-2014 (Semantic Evaluation), which attracted 18 worldwide teams to engage in Aspect category detection aimed at identifying the aspect categories mentioned in a specified review phrase**.**

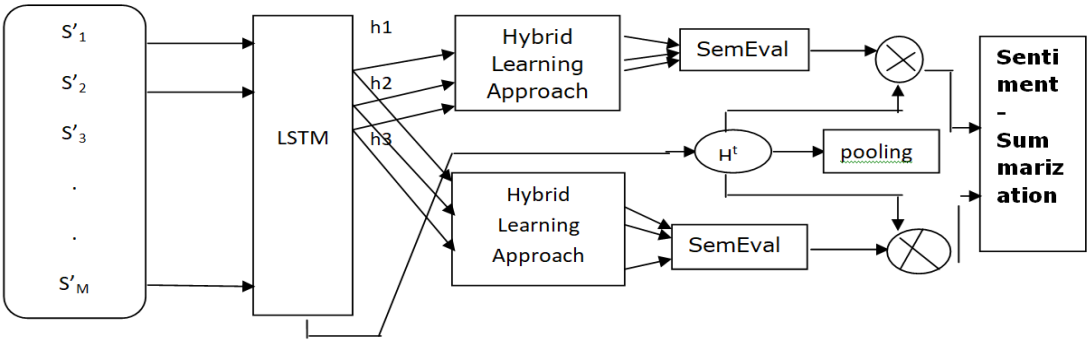

Fig 1. Proposed System Design

Aspect category detection is a significant assessment task for SemEval-2014 (Semantic Evaluation), which attracted 18 teams from around the world to take part. Aspect category detection is intended to define the categories of aspects mentioned in a specified review phrase. Formal, a category of aspect set A={a1,a2, a3... AN} containing N categories is predefined for the domain of the product.For the D = { s1,s2,s3 review dataset SK } containing K phrases, for each phrase we need to predict a binary label vector. Each value in yi shows whether or not the phrase deals with a category of element.In particular, this implies that the phrase si includes the category am aspect and otherwise. The dataset of the restaurant review is used for assessment in SemEval-2014. The domain is predefined by five aspect categories, i.e. A={'food," service," cost," ambience," anecdote / miscellaneous' (brief' a / m')}.

4.Data-set Analysis

We used the dataset of restaurant review released by SemEval-2014 which modified and extended the [4] dataset. There are 3,041 sentences in the training dataset and 800 sentences in the test dataset. Table 1 shows the number of phrases in each category. Fig 2 shows the visual representation of SemEval-2014 data in this blue bar represents testing data and red bar represents testing data.

Category	# of Sentences	
	Training	Test
Food	1232	418
Cost	321	83
service	597	172
ambience	431	118
a/m	1132	234

Table 1: SemEval-2014 Restaurant Review Dataset

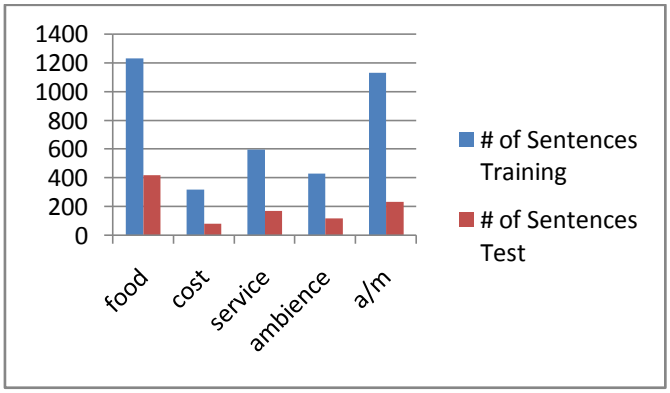

Fig 2: Visual Representation of SemEval-2014 Restaurant Review Dataset

Additionally, we collected an Extended Restaurant Review Dataset to learn the word representations. Part of the dataset is provided by Yelp Dataset Challenge2. The rest of the dataset is crawled from Citysearch3. We use the category names "food", "price", "service" and "ambience/ambiance" as seed words to obtain lots of noisy-labeled sentences **[13]**. Mate-tools is used to parse the dataset and get adjective-noun word pairs via the dependency pattern. Note that the Extended Restaurant Review Dataset is only used for learning better word representations. The final classification model is trained on the SemEval-2014 Restaurant Review Dataset. The detailed statistics of the dataset are shown in Table 2.Fig 3 shows the visual representation of our extended data.

Table 2: Extended Restaurant Review Dataset

# of un-labelled sentences	# of noisy-labelled sentences	# of word pairs
83,24,813	12,14,762	17,90,421

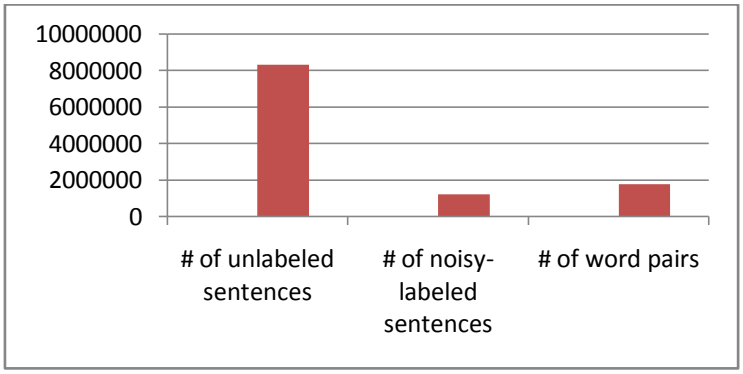

Fig 3: Visual Representation of Extended Restaurant Review Dataset

5. Implementation

We set the vector size as 500 in word representation teaching and the context windows as 5. After word2vec, the learning rate is set at 0.025 and decreases with the training procedure. All word vectors are initialized randomly between -0.5 and 0.5 and the classification weights. We use two 2-layer neural networks to learn the

hybrid characteristics, both of which contain 50 hidden units. We use the marked dataset to train them through back-propagation (not the noisy-labeled dataset). The stochastic descent of the mini-batch gradient is used to update the parameters. The size of the batch is set to 50. The training procedure is run for 500 epochs when the training error becomes steady.

6. Results & Discussion

We use micro F1-score as the assessment metric for all category labels. The most commonly used classification algorithms are LR, NB and SVM. SVM outperforms LR and NB by a big margin between these three techniques. Using unigram and bigram characteristics, it achieves the F1-score over 80.However, the efficiency decreases when the noisy-labeled information are used for practice. It shows that when used directly to train the classifier, the noisy-labeled data cannot enhance the outcome. NRC, which also relies on SVM, is the best system in SemEval-2014. They use an additional lexicon in addition to the textual features, which contains the associations between words and aspects. The lexicon helps from 84.08 to 88.57 boost the performance.The output is even greater than our SVM baseline without the lexicon. That's because their technique uses more complex characteristics like word cluster n-gram. It also demonstrates that feature engineering is a key step towards performance improvement. Four distinct word representation learning algorithms are used here for contrast from the experimental outcomes; we can discover that GloVe and word2vec outperform the other two word embedding algorithms by a big margin on our assignment.Our gathered dataset of restaurant evaluation helps to significantly enhance the efficiency of both word2vec and GloVe. Extracted from word2vec-re, the hybrid characteristics produce similar outcomes with NRC. Our model includes remote monitoring compared to word2vec-re and captures connections between words and elements of sentiment. It achieves the state of the art results with the F1 score of 90.10 on the dataset. Overall, our approach to representation learning outperforms traditional hand-crafted characteristics and embedding algorithms with current words.

7. Conclusion & Futurework

In this research, for aspect category identification, we suggest a hybrid feature learning strategy. We demonstrate that together with the hybrid feature extraction strategy, the Hybrid Aspect Analysis Algorithm provides state-of - the-art efficiency for aspect detection categories. We would like to explore how to enhance the outcomes by injecting internal understanding. We were particularly interested in taking advantage of more semantic options, such as ontology or other semantic networks. We also intend to investigate machine learning methods that solve this issue as we deal with unbalanced data.

References

1. Turney, P. D. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th annual meeting on association for computational linguistics (pp. 417-424). Association for Computational Linguistics.
2. Liu, B. 2011. Opinion mining and sentiment analysis. In Web Data Mining (pp. 459-526). Springer Berlin Heidelberg.Qiu G. 2009. Double Propagation.
3. Ganu, G., Elhadad, N., and Marian, A. 2009. Beyond the Stars: Improving Rating Predictions using Review Text Content. In WebDB.
4. Kiritchenko, S., Zhu, X., Cherry, C., and Mohammad, S. M. 2014. NRC-Canada-2014: Detecting aspects and sentiment in customer reviews. In SemEval 2014, 437.
5. Liu Kang, Xu Liheng, and Zhao Jun. Extracting Opinion Targets and Opinion Words from Online Reviews with Graph Co-ranking. In Proceedings of ACL 2014, Baltimore, USA, June 22-27.
6. Hu, M., and Liu, B. 2004. Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 168-177). ACM.
7. Zhao, W. X., Jiang, J., Yan, H., and Li, X. 2010. Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (pp. 56-65). Association for Computational Linguistics.
8. Mukherjee, A., and Liu, B. 2012. Aspect extraction through semi-supervised modeling. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1 (pp. 339-348). Association for Computational Linguistics.
9. Bengio, Y., Schwenk, H., Senécal, J. S., Morin, F., and Gauvain, J. L. 2006. Neural probabilistic language models. In Innovations in Machine Learning (pp. 137-186). Springer Berlin Heidelberg.
10. Mnih, A., and Hinton, G. E. 2009. A scalable hierarchical distributed language model. In Advances in neural information processing systems (pp. 1081-1088).
11. Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., and Qin, B. 2014a. Learning sentiment-specific word embedding for twitter sentiment classification. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (pp. 1555-1565).

12. Tang, D., Wei, F., Qin, B., Zhou, M., and Liu, T. 2014b. Building Large-Scale Twitter-Specific Sentiment Lexicon: A Representation Learning Approach. In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics (pp. 172-182).
13. Purver, M., & Battersby, S. 2012. Experimenting with distant supervision for emotion classification. In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (pp. 482-491). Association for Computational Linguistics.

