Cross-Domain Sentiment Analysis withDeep Learning

Suman Kumari, Basant Agarwal

Department of Computer Science and Engineering, Swami Keshvanand Institute of Technology, Management and Gramothan, Jaipur-302017(INDIA)

Email- 25suman08@gmail.com, basant@skit.ac.in

Received 23.08.2018 received in revised form 10.09.2018, accepted 12.09.2018

Abstract:**In today's digitized world, reviews and blogs play an important role in getting the knowledge about the best products on the market. These reviews and blogs contain sentiments of peoples related to the products. In the field of natural language processing, sentiments are classified under the area of sentiment analysis and have always been the active area of research. The labeled data is considered as source domain and is trained for sentiment classification. The sentiment value change from domain to domain due to which when the classification is performed on cross-domain dataset, it produces the result differently. In this paper, we presented the cross-domain semantic analysis on product review dataset by capturing the semantic feature with word embedding algorithm and processing it with deep learning neural network to improve accuracy for cross-domain sentiment classification tasks.**

Keywords:Cross-domain, Word2vec, Sentiment classification, Conventional Neural Network

2. INTRODUCTION

The rapid growth of the internet has made the user express their views online about products they buy. These views are the sentiments related to the products. Estimating these sentiments is necessary as they provide necessary information about the product. If the sentiment is negative, it informs the seller that the product has some problem and if the sentiment is positive, it is helpful for the consumer. The field under natural language processing that deal with sentiments of human text information is sentiment analysis. It helps a customer to easily understand the overall sentiments related to the product by automatically classifying the reviews of another customer about the product. To perform sentiment analysis various models are used but using deep learning model has shown its effectiveness. For the process of sentiment classification training and testing model for same domain or different domain is considered. Sentiment analysis becomes challenging when training and testing data is from a different domain. This is because, sentiments behavior to products changes from one domain to another and also there are many such words which do not have a connection with other domain.

The NLP system has word as its basic units for input and output. For sentiment classification, a vector representation is generated to catch semantic and syntactic connection of natural language [1]. In natural language processing, there are two particular methods for encoding word into a vector representation. These methods are: One hot encoding and Word embedding.

1.1 One –hot encoding

The traditionally used method was one hot encoding that transforms a word into the form of the vector representation but only has two values 0 (if the word does not exist) or 1 (if the word exist).

1.2 ***Word embedding***

Word embedding is a non-twofold word representation model which frame words from a vocabulary into a vector representation. It considers the co-occurrenceproperties of the words and arranges words according to its syntactic and semantic relation. It is nowadays most popular technique for representation of words in the natural language processing system. The term was stated in 2003 by bengio et. al. [2]. Figure 1 shows the vector representation of text with the word embedding algorithm.

Word embedding is a non- two fold word representation model which frame words from vocabulary into vector representation. It considers the co-occurrence properties of the words and arranges words according to its syntactic and semantic relation. It is now day‟s most popular technique for representation of words in natural language processing system. The term was stated in 2003 by bengio et al.[2].

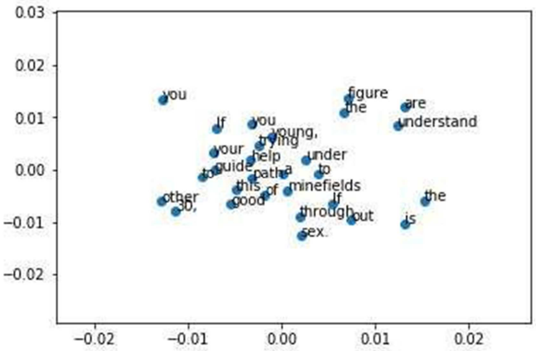

Figure 1: Vector representation of text with word embedding algorithm.

Word embedding uses various algorithms to learn representation for the words one of them is word2vec [3]. Word2vec take a corpus of the word as input and produces vector as its output. The training dataset is processed first to form vocabulary and then vector representation of the word is learned. The texts processed by word2vec are handled by the deep learning algorithms. Deep learning [4] has the ability to learn vector representation. It is referred to as deep learning network. It is used to provide training for both supervised and unsupervised learning of datasets. It includes many networks such as convolutional neural network, recurrent neural network and some more. Convolutional neural network [5] has less number of parameters and is easy to be trained.

2. RELATED RESEARCH

Sentiment analysis has been proven one of the most vital tools in the recent years. The world is moving towards digitization. Most of the work is being carried out digitally. In the field of sentiment analysis, word embedding plays a vital role. Lots of work has been done for sentiment analysis but workings with the cross-domain dataset are still a few. The work related to Sentiment analysis with the cross-domain dataset is mentioned in Table 1.

3. PROPOSED MODEL

The proposed model includes pre-processing as the initial stage. Pre-processing provides a clean and decent representation of the document. Then the processed document is provided for vector representations of the word which is done by word embedding algorithm. Then the sentiment classifier method is used for evaluating the method. The deep learning neural network classifier provides high-quality classification so the whole process is processed with convolution neural network. The trained model is evaluated on the various cross-domain dataset. Proposed approach is depicted in figure 2.

3.1 Pre-processing Step

First of all, the data is preprocessed to remove the noisy and irrelevant data. We use tokenization, stop word removal, stemming to remove the noisy features.

3.1.1 Tokenization

Tokenization is the process of splitting the sentences in the form of tokens and to identify space, comma, and special symbols. Words are separated by the comma delimiter in the proposed approach.

3.1.2 Stop Word Removal

Stop words are the commonly used words such as articles which do not have their own meaning. These stop words must be removed from the tokenized words to get efficient word representation. There is some predefined list of English stop words some of these are shown in table 2 shows the result of stop word removal step when applied to tokenized words.

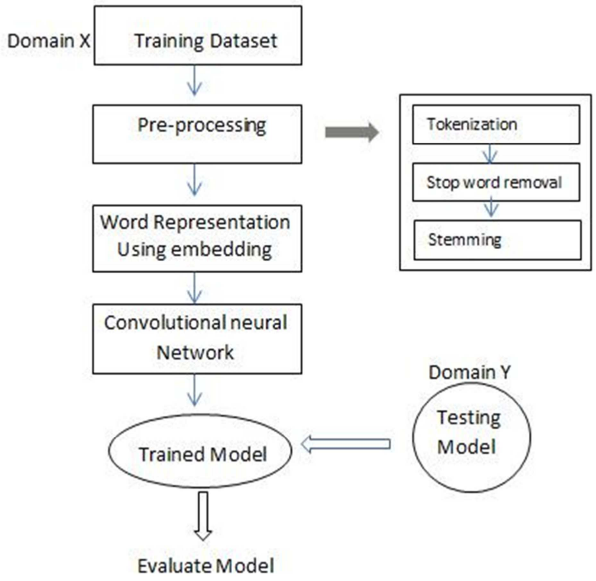

Figure 2: Proposed Approach

3.1.3 Stemming

Stemming is the way toward distinguishing stem or root words to get a typical base word for the words. The words which are syntactically similar in nature such as plural words and word with verbal variations etc. all these are considered as same words. For example words like started, starting, starts are considered similar as they as derived from single word start.

Table 1: Related work

Paper	Work	Year
Shai Ben-David [6]	Distance between domains as measure of the loss due to reworking from one to the other.	2006
John Blitzer [7]	Domain adaption algorithm for classifiers on reviews for cross domain products	2007
Tao Li [8]	Cross domain sentimental analysis using standard text categorization methods.	2009
Niklas Jakob [9]	Opinion mining with cross domain dataset based on conditional random field.	2010
Jialin Pan [10]	Cross domain sentiment classification via spectral feature alignment and compared their method with SCL	2010
Yulan He [11]	Modifying joint sentiment topic model with multi domain sentiment dataset.	2011
Rui Xia and Chengqing Zong [12]	Sentiment classification with cross domain featured with part-of-speech tags.	2011
K Paramesha [13]	Enhancing performance for Sentimental analysis for cross domain dataset using Sentiwordnet	2013
D. Bollegala [14]	Thesaurus with cross domain dataset to classify sentiments related to user review.	2013
N. Kalchbrenner [15]	Dynamic Convolutional neural network for semantic modelling of sentences.	2014
Mauro Dragoni and Giulio Petrucci [16]	NeuroSent tool for enabling the building of multi-domain sentiment model.	2017
Dirk von Grunigen [17]	Performance analysis with cross domain dataset trained on convolutional neural network.	2017
Shai Ben-David [6]	Distance between domains as measure of the loss due to reworking from one to the other.	2006
John Blitzer [7]	Domain adaption algorithm for classifiers on reviews for cross domain products	2007
Tao Li [8]	Cross domain sentimental analysis using standard text categorization methods.	2009
Niklas Jakob [9]	Opinion mining with cross domain dataset based on conditional random field.	2010
Jialin Pan [10]	Cross domain sentiment classification via spectral feature alignment and compared their method with SCL	2010
Yulan He [11]	Modifying joint sentiment topic model with multi domain sentiment dataset.	2011
Rui Xia and Chengqing Zong [12]	Sentiment classification with cross domain featured with part-of-speech tags.	2011
K Paramesha [13]	Enhancing performance for Sentimental analysis for cross domain dataset using Sentiwordnet	2013
D. Bollegala [14]	Thesaurus with cross domain dataset to classify sentiments related to user review.	2013
N. Kalchbrenner [15]	Dynamic Convolutional neural network for semantic modelling of sentences.	2014
Mauro Dragoni and Giulio Petrucci [16]	NeuroSent tool for enabling the building of multi-domain sentiment model.	2017
Dirk von Grunigen [17]	Performance analysis with cross domain dataset trained on convolutional neural network.	2017

3.2 Word2Vec

Word embedding is performed to accomplish word representation in vector space. Embedding's are generally prepared either by training the data or by pre-trained data. Word2vec algorithm is used to form effective embedding by training the data. Word2vec algorithms form a predictive model with low dimensional vector space representation in which words with the semantic relationship are close with cosine distance. Word2vec model has two methods for performing representation. The methods are CBOW (Continuous bag of the word) and skip gram. Both methods work inverse of each other as Continuous bag of the word model predict word from the context while skip gram predicts context from the word. It works with one of themethods. In this paper, text corpus of training dataset is processed with a word2vec algorithm with skip gram model to get a vocabulary of words and to form effective word representation model for words.

Table 2: Stop Word List: Some Example

They	For	Then
Those	Over	Been
Whom	is	Such
This	more	Because
About	That	Other
Itself	Herself	Before
Are	Any	And
But	By	Can
Both	Between	Being
Myself	Nor	More
Most	Your	Did

3.3 Convolutional Neural Network

The deep learning neural network method is used to get a high-quality representation of words. One of the deep learning networks is a convolutional neural network which is efficient for learning features and is being used in many areas such as spam detection [18], paraphrase detection [19], image reorganization and many others. Convolution one-dimensional neural network is used in this paper which takes input as 1-D word embedding matrix and output a trained classifier model. Convolution neural network is a multi-neural network with convolutional and pooling layer that goes with the fully connected layer. It also contains a convolutional filter which performs dot products and helps in the feature extraction process. The convolution network contains less number of parameter and some of them are optimizers, learning rate, and dropout. Changing the value of each parameter changes the result.

4. EXPERIMENT AND RESULT

The experimentsare performed on anaconda cloud which combines libraries for python and also the IDE with IPython.

4.1 Dataset

The product review cross domain dataset is considered for an experiment which contains a review of four different domain i.e. DVD, Book, electronic, and kitchen. The sets of the cross-domain dataset are formed to evaluate the method. Datasets are represented in terms of a letter such as a book is B, Dvd is D, electronics is E and the last set kitchen is K. Considering one of the datasets as a training dataset and other as a testing dataset evaluation is performed for finding accuracy.

4.2 Result

To produce the result and to get the effectiveness of system experiment is performed. When we deal with convolution neural network there is a number of factors that affect the performance such as its optimizer, numbers of hidden layer, increase in dropdown value and many others. The experiment is performed for analyzing accuracy proposed model when applied with the cross-domain dataset. Table 3 shows the effect.

Table 3: Accuracy with cross domain dataset

Datasets ↓	Word2Vec
D-E	69.2
K-E	70.6
E-D	63.4
D-K	66.1
B-D	71
E-B	59.35
E-K	74.4
K-B	59.8

The combination of dataset generally performs well with a dropout of 0.4, and with 3 hidden layers. The system is evaluated with various optimizers also and it shows the different result as shown in figure 3.

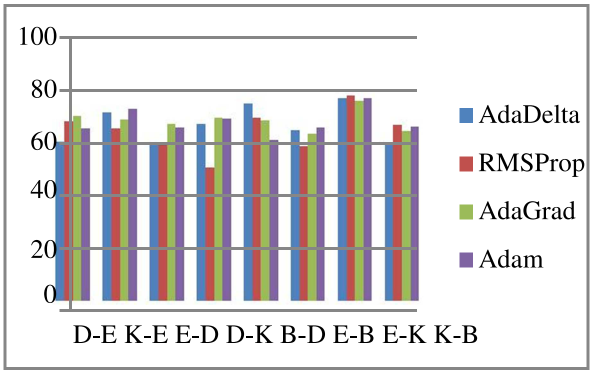

Figure 3: Dataset with different optimizers

5. CONCLUSION AND FUTURE WORK

Natural language processing contain a field in which we can automatically classify sentiments related to text information, this field is named as sentiment analysis.

Analyzingsentiments related to texts are useful, as nowadays people like to share their opinion about the product online and it is the only way to analyse the product from online stores. The rapid growth of internet has made the sentiment analyser more effective. In this paper, we applied a one-dimension convolutional neural network algorithm for sentiment classification which takes input as vector representation of word processed by the word2vec algorithm. The whole model is trained on a dataset of various domains and the dataset of different domains are used as testing dataset. The cross-domain sentiment analysis is performed to evaluate the accuracy of themodel. In Future, this work can be extended to improve the performance of the model by applying some methods as due to cross-domain dataset performance of the system is low.

REFERENCES

[1] Mikolov, T., Chen, K., Corrado, G., Dean, J.,“ Efficient estimation of word representations in vector space”, In: Proceedings of international conference on learning representations, 2013.

[2] Bengio, Y., Ducharme, R., Vincent, P,.“Neural Probabilistic Language Model”, Journal of Machine Learning Research, pp: 1137–1155, 2003.

[3] Mikolov, T., Chen, K., Corrado, G., and Dean, J., “Efficient estimation of word representations in vector space”, ICLR Workshop, 2013.

[4] Hinton, G., E., and Salakhutdinov, R., R.,“Reducing the dimensionality of data with neural networks”, pp. 504-507, 2006.

[5] Yoon Kim, “Convolutional Neural Networks for Sentence Classification”, arXiv:1408.5882v2 [cs.CL], 3 Sep 2014.

[6] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira, Analysis of representations for domain adaptation. In Neural Information Processing Systems (NIPS), 2006.

[7] John Blitzer, Mark Dredze, and Fernando Pereira, “Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification”, In Proceedings of Association for Computational Linguistics, pp. 440– 447, 2007.

[8] Tao Li, Vikas Sindhwani, Chris Ding and Yi Zhang, “Knowledge transformation for cross-domain sentiment classification”,

[9] Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval , 2009.

[10] Niklas Jakob and Iryna Gurevych, “Extracting Opinion Targets in a Single- and Cross-Domain Setting with Conditional Random Fields”, 2010.

[11] Jialin Pan, Xiaochuan Ni, Jiantao Sun, Qiang Yang and Zheng Chen, Cross-domain sentiment classification via spectral feature alignment. In Proceedings of the International World Wide Web Conference (WWW), pages 751-760, 2010.

[12] Yulan He, Chenghua Lin and Harith Alani, “Automatically Extracting Polarity-Bearing Topics for Cross-Domain Sentiment Classification”, Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, Pages 123-131, 2011.

[13] Rui Xia and Chengqing Zong, “A POS-based Ensemble Model for Cross-domain Sentiment Classification”, 2011.

[14] K Paramesha and K C Ravishankar, “optimization of cross domain sentiment analysis using sentiwordnet”, 2013.

[15] Danushka Bollegala, David Weir, and John Carroll, “Cross-domain sentiment classification using a sentiment sensitive thesaurus”, IEEE Transactions on Knowledge and Data Engineering, pp.1719–1731, 2013.

[16] Kalchbrenner N, Grefenstette E, Blunsom P, “A convolutional neural network for modelling sentences,” arXiv preprint arXiv:1404.2188, 2014.

[17] Mauro Dragoni and Giulio Petrucci, “A Neural Word Embeddings Approach for Multi-Domain Sentiment Analysis”, 2017.

[18] Weilenmann, Martin & Deriu, Jan & Cieliebak, Mark & vonGrünigen, Dirk, “Potential and Limitations of Cross-Domain Sentiment Classification”, 10.18653/v1/W17-1103, 2017.

[19] Gauri Jain, Manisha, Basant Agarwal, “Spam Detection on Social Media using Semantic Convolutional Neural Network”, In International Journal of Knowledge Discovery in Bioinformatics (IJKDB), IGI Global, Vol 8 (1), pp: 12-26, 2018.

[20] Basant Agarwal, Heri Ramampiaro, Helge Langseth, Massimiliano Ruocco , “A Deep Network Model for Paraphrase Detection in Short Text Messages”, Information Processing and Management, Elsevier, 54(6), Pages 922-937, 2018.