INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

An Approach for Aspect Extraction Using Double Embedding Technique based Machine Learning Model

Divya Jatain[1]*, **Vikram Singh**[2]

1 Department of Computer Science & Engineering, Maharaja Surajmal Institute of Technology, New Delhi, India
2 Department of Computer Science & Engineering, Chaudhary Devi Lal University, Sirsa, India

OPEN ACCESS

Received: 10-01-2023
Accepted: 13-03-2023
Published: 12-05-2023

Citation: Jatain D, Singh V (2023) An Approach for Aspect Extraction Using Double Embedding Technique based Machine Learning Model. Indian Journal of Science and Technology 16(19): 1408-1412. https://doi.org/10.17485/IJST/v16i19.62

* **Corresponding author**.

divyajatain@msit.in

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: To perform aspect extraction for opinion mining of learner comments in the online teaching-learning scenario. **Methods :** A machine learning model is developed for aspect extraction. The authors collected the dataset consisting of around 5000 learner comments from Coursera and performed analysis on the dataset. To validate the results, the standard SemEval2014 dataset is used. **Findings:** For both the contextualised and non contextualised word embeddings, the authors compiled the results of their model for two different datasets, viz., the SemEval 2014 dataset and the collected dataset. The proposed model provided an accuracy of 92.78%. **Novelty:** For the first time in the literature, double embedding is used for aspect extraction. Using domain-specific embedding provided good results, but the proposed model (using double embedding) is far superior in accuracy.

Keywords: Opinion Mining; Feature Extraction; Double Embedding; Learner Comments; Contextualised Embeddings

1 Introduction

The technological revolution, along with the technological enhancements in the Internet and mobile devices, has made people spend a significant amount of their time in an online world where every person is a content consumer and content creator. The data that is generated by the people can be used for making a multitude of decisions in varied domains like marketing & research, advertising, generating specialised recommendations, etc. Opinion mining can be understood as an intersection of different techniques involving computer science, linguistics, and statistics to identify the opinion of the majority of people towards some topic (generally a product, service, or some issue) in the discussion. Thus, mining people's opinions is a critical task to identify, analyse and utilise insights for critical decision-making. When it comes to influencing individuals and their beliefs, people with a larger digital footprint can also be strategically manipulated to shape opinions on subjects with a wide range of interests and connections to many fields.

In 2011, a revolution has started in the field of teaching-learning in the form of Massive Open Online Courses (MOOCs) [1–3]. Ever since, the domain of online

teaching-learning continues to grow at high speed and online teaching-learning has been an important knowledge disseminator in the corona and post-corona world. Identifying and analysing the opinions & emotions of learners towards some topic, subject taught, course as a whole, or the instructor can not only help the course-content creators design efficient & well-liked courses which are popular among the students but also improve the entire experience. Thereby, making the teaching-learning activity fruitful.

According to research work, in the past decade, there has been an increasing trend of application of machine learning, artificial intelligence, and natural language processing for mining opinions in the education domain [4]. Some works have been carried out to identify the polarity of a learner feedback comment. Still, the limitation is that they need to consider the overall aspect related to the polarity. For instance, consider the feedback given by a learner "The explanation provided is very lucid but wish the pace was faster." Here, the sentence polarity is positive with respect to the aspect explanation but not so with respect to the aspect pace of teaching. For fine grained opinion mining, the analysis should be performed on the aspect terms rather than the review sentence as a whole.

For performing some useful analysis, the review comments first need to be effectively converted into machine-readable number form (the process is called vectorisation). The methods like Bag of words or tf*idf are found to be having specific issues. These processes were computationally expensive; there needed to be a more meaningful relationship between the words and no consideration for the word order. So, word embeddings came into the picture.

Word embeddings are the representation of words into real-valued vectors for text analysis. There are non-contextualised embedding methods like FastText [5], GloVe, Word2Vec. But to have a better, fine-grained detail about the aspect, it is important to have domain specific knowledge also. So, the authors in this work propose a double embedding method for a deep learning RNN model to perform aspect-based opinion mining. The results of the model with the proposed double embedding is compared with the one with contextualised embedding and the one with domain-specific embedding only for the MOOC Dataset (a dataset collected by the authors having around 5000 comments fetched from the course videos uploaded over Coursera) and for the standard SemEval 2014 task 4 dataset. Experimental results show the superiority of the proposed method over the others by an accuracy of 92.78%.

The aim of opinion mining & sentiment analysis is to identify the opinion of people with respect to some object, product, services, or some topic in discussion. People's opinion may either be positive, negative, or neutral, where the positives and negatives can further be identified to different varying degrees or intensities. Opinion mining can be used for the identification of customers' opinions [6,7] adopting and formulating better marketing strategies [8], building better recommendation systems, developing better healthcare services based on patient's moods and opinions towards disease, drug reactions, etc [9]. With the development of new technologies viz., Cloud Computing, Blockchain [10] and Big Data, a whole new world of opinion mining being applied to varied fields has opened up.

From the perspective of Artificial Intelligence and Machine Learning, opinion mining can be performed at the level of document, sentence and aspect. Extraction of Aspects is an important task in Natural Language Processing and has a wide range of application domains [11,12]. For instance, considering review statements in the Persian language [13], for sentiment analysis of tweets related to Covid-19 [14], for hotel reviews [15], for online product reviews [6], etc. Effectively, aspect extraction can be performed using unsupervised approaches like frequent pattern mining [16], syntactic rules-based extraction, topic modelling [17], word alignment and label propagation or supervised approaches [18], Conditional Random Fields (CRF). Deep neural networks can also be applied for aspect extraction, e.g., using LSTM, attention mechanism [19], co-extraction via a deep network, etc.

As far as the education domain is concerned, few researchers have worked. An interesting research that uses a lexicon-based approach has used Pearson correlation for the task of identifying document-level polarity. Lexicon based approach has also been used to provide details regarding the teacher's performance using a metric similar to the Likert Scale. Research work has also been carried out focusing combination of lexicon-based features with machine learning. For improving the overall teaching process & the early prediction of student performance is a key factor, Yu et al. have done a sentiment analysis of self-evaluated student comments [20].

In order to improve teaching-learning process, Chauhan et al., have used Stanford NLP parser is used to extract aspect information and perform aspect based opinion mining [21]. Another significant research involves the use of weak supervision signal of the Dataset specific aspects for identification of aspect category and further aspect polarity [22]. In the next section we discuss the preliminaries of word embeddings so as to form a background for our dual embedding method.

Opinion mining is a subtask of Natural Language Processing wherein the review comments effectively need to be converted into numbers in order to be understood & processed by machines. This process of conversion of textual human-readable reviews into machine-readable numbers is known as Vectorisation.

Word embeddings are the most commonly used vectorisation techniques[23,24]. Most of the research works in the field of aspect extraction use pre-trained embeddings like FastText[5], GloVe, Word2Vec. However, in order to get a fine-grained detailed meaning, domain-specific embeddings are very crucial. So, in this work, the authors propose a dual embedding-based Recurrent Neural Network for aspect extraction. The choice of a Recurrent Neural Network is based on Occam's Razor principle which states that a simpler and easier method is always preferred over a complex method. In the next section, the authors will discuss the details of the proposed dual embedding-based model.

1.1 Gaps in the Literature & the Contributions of the Work

Though there are a few works dealing with opinion mining in the education domain, there is still scope for improvement because, majorly, the aspect extraction is carried out using either contextualised or non-contextualised embeddings. No known research work has used double embedding for performing the task of aspect extraction, and this makes the work unique.

Moreover, the authors have done a comparative analysis of some recent approaches to aspect extraction with their proposed double embedding method, as shown in Figure 1, for a standard dataset to avoid bias. The results show the superiority of the proposed model over the existing works.

2 Methodology

The architecture proposed in this work consists of two layers of embeddings viz., domain specific and general purpose embedding, followed by a four layered Recurrent Neural Network architecture. There are a total 7 layers, with 2 embedding and 1 convolution layer. The activation function used is softmax and the fully connected layer. The loss function used in the model is sparse categorical cross entropy.The fully connected layer is shared across all the layers to identify the positions of words.

For the working model, the input sequence of the review comments can be represented as a set as R= (r_1, r_2, r_3,….,r_n). The input review sequence can be represented by successive double embeddings in the form of R_g (general non-contextualised embedding, E_g) and R_d (domain specific embedding, E_d which is trained over the dataset collected by the authors). Further, these embeddings are combined to form a dual embedding as follows

$$E' = E_g \oplus E_d$$

and fed into the further layers of the model. In the proposed model, the convolution operations and ReLu activations are performed according to the following equation

$$R_{i,j}^{k+1} = max(0, (\sum_{n=-c}^{c} E_{n,j}^{k} R_{i+n}^{k}) + b_j^k)$$

where k is the k^{th} layer of the model. Filter is applied to the positions from 1 to n such that each filter computes the representation of the word and thus, the output of the layers becomes aligned with the original input R.Other details of the architecture such as the output shape, parameters- total, trainable and non-trainable are mentioned in Table 1 below.

Table 1. Summary of the Architecture of the Proposed Model

Layer (type)	Output Shape	param #
Embedding_1 (DomainSpecificEmbedding)	(None, 156, 50)	5000
Embedding_2 (GeneralEmbedding)	(None, 156, 50)	5000
batch_normalization (BatchNo)	(None, 156, 50)	200
spatial_dropout1d (SpatialDr)	(None, 156, 50)	0
rnn (SimpleRNN)	(None, 156, 100)	640
batch_normalization_1(Batch)	(None, 156, 100)	40
dense_1 (Dense)	(None, 100)	804
dense (Dense)	(None, 8)	80
Total params : 14,044 Trainable params : 14,044 Non-trainable params: 0		

For the proper functioning of the model, dropout is applied to the embedding layer and after each ReLU activation. The given model is run for the different non- contextualised embeddings and for the dataset created by the authors and the standard SemEval2014 dataset. The experimental findings are discussed in the next section.

3 Results and Discussion

The authors have discussed the detailed experimental results for the aspect extraction task in this section. For this, in the first set of experiments, the authors have considered the dataset collected and created by them, which consists of around 5K learner comments from Coursera.

The authors have thoroughly evaluated the performance of their model by using non-contextualised embeddings and then using a concatenation of non-contextualised word embeddings with domain-specific embedding (created by the authors) for the proposed model. The corresponding results are represented in Table 2. The model that uses GloVe along with domain-specific embedding provides better results for the parameters of Precision and F1 score.

Table 2. Summary of Results of Aspect Extraction for Educational Dataset (Collected from Coursera)

Model	F1 score	Precision	Recall
Domain Specific+RNN	80.70	82.10	80.14
FastText+RNN	78.84	81.16	81.37
GloVe+RNN	79.75	80.10	80.87
Word2Vec+RNN	81.75	82.10	81.59
(FastText+ Domain Specific) +RNN	81.59	81.89	81.99
(GloVe+ Domain Specific) +RNN	82.13	83.31	81.85
(Word2Vec+ Domain Specific) +RNN	81.28	81.61	81.95

For performing a comparative analysis, the authors have used a standard dataset. This is because of the fact that for a self-collected dataset, there may be a situation of bias. So, to avoid this, SemEval 2014 dataset, which is a well-known dataset is used to compare the proposed model with some of the pre-existing models present in the literature. For this, the input parameters are slightly changed. The SemEval-14 Dataset has five labelled aspects and two polarity orientations. So, output parameters are changed from four to five, and only positive and negative opinion orientations are considered. Also, some other research works that have used aspect extraction for sentiment analysis are considered and their corresponding results for the standard SemEval-14 Dataset are represented in Figure 1.

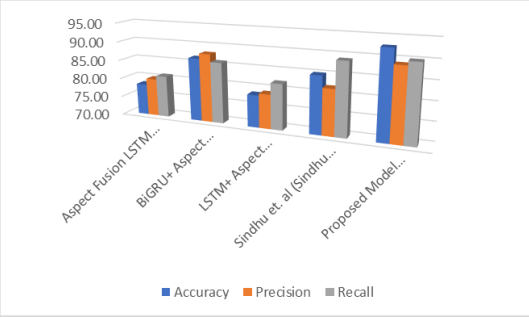

Fig 1. Comparison of Proposed Model with state-of-art approaches for Sentiment Polarity Identification for SemEval 2014 Dataset

4 Conclusion

The model proposed in this work follows Occam's razor Principle by being simple, and its efficiency in terms of extraction of aspects is demonstrated experimentally by the authors in the previous section. The results for both the datasets, i.e., for the dataset collected by the authors and from the standard SemEval2014, show the superiority of the proposed model. There are a few studies in the literature related to aspect extraction in the domain of education, but in the current scenario, the studies have used either contextualised or non-contextualised word embeddings. For the first time, a dual embedding model is presented, which makes this work novel.

The current work can be limited by the fact that absence of some standard dataset in the education domain makes the model subjective to varied interpretation. In the future, this work can be extended by analyzing the sarcastic comments, emoji inclusion, and having more cues from the informal communication manner.

Acknowledgements

The authors would like to thank Maharaja Surajmal Institute of Technology, New Delhi and Department of Computer Science & Engineering, Chaudhary Devi Lal University, Sirsa for providing help at all stages of this research work.

References

1) Gardner J, Brooks C. Student success prediction in MOOCs. *User Modeling and User-Adapted Interaction*. 2018;28(2):127–203. Available from: https://doi.org/10.1007/s11257-018-9203-z.
2) Deng R, Benckendorff P, Gannaway D. Progress and new directions for teaching and learning in MOOCs. *Computers & Education*. 2019;129:48–60. Available from: https://doi.org/10.1016/j.compedu.2018.10.019.
3) Zhu M, Sari A, Bonk CJ. A Systematic Review of MOOC Research Methods and Topics : Comparing. *Proceedings of EdMedia: World Conference on Educational Media and Technology*. 2013. Available from: https://www.learntechlib.org/primary/p/184395/.
4) Kastrati Z, Dalipi F, Imran AS, Nuci KP, Wani MA. Sentiment Analysis of Students' Feedback with NLP and Deep Learning: A Systematic Mapping Study. *Applied Sciences*. 2021;11(9):3986. Available from: https://doi.org/10.3390/app11093986.
5) Mikolov T, Grave E, Bojanowski P, Puhrsch C, Joulin A. Advances in pre-training distributed word representations. *Computation and Language (csCL)arXiv*. 2019;p. 52–57. Available from: https://doi.org/10.48550/arXiv.1712.09405.
6) Zhao H, Liu Z, Yao X, Yang Q. A machine learning-based sentiment analysis of online product reviews with a novel term weighting and feature selection approach. *Information Processing and Management*. 2021;58(5):102656. Available from: https://doi.org/10.1016/j.ipm.2021.102656.
7) Kumar S, Yadava M, Roy PP. Fusion of EEG response and sentiment analysis of products review to predict customer satisfaction. *Information Fusion*. 2019;52:41–52. Available from: https://doi.org/10.1016/j.inffus.2018.11.001.
8) Bernabé-Moreno J, Tejeda-Lorente A, Herce-Zelaya J, Porcel C, Herrera-Viedma E. A context-aware embeddings supported method to extract a fuzzy sentiment polarity dictionary. *Knowledge-Based Systems*. 2020;190:105236. Available from: https://doi.org/10.1016/j.knosys.2019.105236.
9) Ramírez-Tinoco FJ, Alor-Hernández G, Sánchez-Cervantes JL, Salas-Zárate MDP, Valencia-García R. Use of Sentiment Analysis Techniques in Healthcare Domain. In: G AH, JL SC, A RG, R VG, editors. Studies in Computational Intelligence. Springer International Publishing. 2019;p. 189–212. Available from: https://doi.org/10.1007/978-3-030-06149-4_8.
10) Frizzo-Barker J, Chow-White PA, Adams PR, Mentanko J, Ha D, Green SE. Blockchain as a disruptive technology for business: A systematic review. *International Journal of Information Management*. 2020;51:102029. Available from: https://doi.org/10.1016/j.ijinfomgt.2019.10.014.
11) Jatain D, Singh V, Dahiya N. A multi-perspective micro-analysis of popularity trend dynamics for user-generated content. *Social Network Analysis and Mining*. 2022;12(1):147. Available from: https://doi.org/10.1007/s13278-022-00969-7.
12) Divya, Singh V, Dahiya N. Computational Intelligence Techniques for Big Data Analytics: A Contemplative Perspective. In: Lecture Notes in Electrical Engineering;vol. 2022. Springer Singapore. 2022;p. 391–400. Available from: https://doi.org/10.1007/978-981-16-8248-3_32.
13) Vazan M, Masoumi FS, Majd SS. Joint Learning for Aspect and Polarity Classification in Persian Reviews Using Multi-Task Deep Learning. 2022. Available from: https://doi.org/10.48550/arXiv.2201.06313.
14) Basiri ME, Nemati S, Abdar M, Asadi S, Acharrya UR. A novel fusion-based deep learning model for sentiment analysis of COVID-19 tweets. *Knowledge-Based Systems*. 2021;228:107242. Available from: https://doi.org/10.1016/j.knosys.2021.107242.
15) Muhammad PF, Kusumaningrum R, Wibowo A. Sentiment Analysis Using Word2vec And Long Short-Term Memory (LSTM) For Indonesian Hotel Reviews. *Procedia Computer Science*. 2021;179:728–735. Available from: https://doi.org/10.1016/j.procs.2021.01.061.
16) Xu H, Liu B, Shu L, Yu PS. Lifelong Domain Word Embedding via Meta-Learning. *arXiv*. 2018. Available from: https://doi.org/10.48550/arXiv.1805.09991.
17) Han Y, Moghaddam M. Analysis of sentiment expressions for user-centered design. *Expert Systems with Applications*. 2021;171:114604. Available from: https://doi.org/10.1016/j.eswa.2021.114604.
18) Hu M, Peng Y, Huang Z, Li D, Lv Y. Open-Domain Targeted Sentiment Analysis via Span-Based Extraction and Classification. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019. Available from: https://aclanthology.org/P19-1051.
19) Liu M, Li L, Hu H, Guan W, Tian J. Image caption generation with dual attention mechanism. *Information Processing & Management*. 2020;57(2):102178. Available from: https://doi.org/10.1016/j.ipm.2019.102178.
20) Yu LC, Lee CW, Pan HI, Chou CY, Chao PY, Chen ZH, et al. Improving early prediction of academic failure using sentiment analysis on self-evaluated comments. *Journal of Computer Assisted Learning*. 2018;34(4):358–365. Available from: https://doi.org/10.1111/jcal.12247.
21) Chauhan GS, Agrawal P, Meena YK. Aspect-Based Sentiment Analysis of Students' Feedback to Improve Teaching–Learning Process. *Information and Communication Technology for Intelligent Systems*. 2019;2:259–266. Available from: https://doi.org/10.1007/978-981-13-1747-7_25.
22) Kastrati Z, Imran AS, Kurti A. Weakly Supervised Framework for Aspect-Based Sentiment Analysis on Students' Reviews of MOOCs. *IEEE Access*. 2020;8:106799–106810. Available from: https://doi.org/10.1109/ACCESS.2020.3000739.
23) Hew KF, Hu X, Qiao C, Tang Y. What predicts student satisfaction with MOOCs: A gradient boosting trees supervised machine learning and sentiment analysis approach. *Computers & Education*. 2020;145:103724. Available from: https://doi.org/10.1016/j.compedu.2019.103724.
24) Dessì D, Dragoni M, Fenu G, Marras M, Recupero R, D. Deep Learning Adaptation with Word Embeddings for Sentiment Analysis on Online Course Reviews. 2020. Available from: https://doi.org/10.1007/978-981-15-1216-2_3.