Neural Network based Approaches for Aspect-Based Sentiment Analysis

Qingjie Lu *

Department of Computer Science, University of Rochester, Rochester NY 14623, USA

* Corresponding Author Email: qlu7@u.rochester.edu

Abstract. The research of Aspect-based Sentiment Analysis which is a process that has a more specific focus than general sentiment analysis is trending upwards in numbers. Stemming from Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs), novel approaches introduced new components like Graph Convolutional Networks (GCNs) and Transformers that improved the overall accuracy dramatically. Along with summarizing the models, the focus of this survey will be on comparing the several novel methods. Although this paper found that Dependency graph enhanced dual-transformer network (DGEDT) coupled with Bidirectional Encoder Representations from Transformers (BERT) is the best performing model thus far, this paper also identified challenges that needed to be addressed in order to better evaluate current and future models.

Keywords: natural language processing; aspect-based; sentiment analysis; novel methods

1. Introduction

Aspect-based sentiment analysis (ABSA) is a branch of Natural language Processing (NLP) that aims to extract a more detailed level of information. Rather than learning the general sentiments of a sentence, the purpose of ABSA is to identify the aspect expression and the corresponding sentiment expression of certain sentences. For example, in the sentence "this paper is good," "paper" should be recognized as the aspect expression, and the sentiment expression or "polarity" is "good."

ABSA is of high importance as it can provide huge benefits to business applications [11&16]. For instance, ABSA can help businesses process huge amounts of unstructured data in an efficient and cost-effective way, which will give businesses more insight feedbacks on the consumer needs and help businesses improve their products and services. ABSA can also enable new possibilities in terms of real-time analytics. For example, Public Relations companies can use ABSA to identify critical issues in real-time so that they may take action immediately.

Early attempts of ABSA consist of direct applications of deep learning algorithms: Recurrent Neural Network (RNN) with Long-short-term-memory (LSTM) or Gated Recurrent Unit (GRU) can process sequential data and capture long term dependencies and structures; Convolutional Neural Network (CNN) can extract local representation of text [13]. Attention mechanism that can help the model focus on the important parts of the sentences has been introduced into the neural networks as well which does improve the performance and shows promises[14].

While the early attempts demonstrate good accuracy, more recent models have demonstrated significant improvements over baseline models. New architectures such as Graph Convolutional Network and Dual Transformers have been introduced to address the shortcomings of early models. Among all the models that this paper examined thus far, Dependency graph enhanced dual-transformer network (DGEDT) coupled with Bidirectional Encoder Representations from Transformers (BERT) showed the best performance overall against a range of other baseline and sophisticated models.

This paper seeks to explore and summarize a couple of the novel ABSA methods: introducing components of models, analyzing the ablation studies and case studies. This paper will examine the results of the several methods and evaluate their performances together to identify the best performing models. This paper will also discuss future investigations and challenges that should be done to better evaluate the effectiveness of each model.

2. Related research

This section details the architectures of each model and identifies the critical components that drive the improvements in accuracy.

2.1. Aspect Specific Graph Convolutional Network (ASGCN)

ASGCN was proposed to address the challenges of using unrelated context words as descriptors and determining the aspects of non-consecutive words [1]. The architecture of the ASGCN is as follows: The network starts with a traditional Recurrent Neural Network with LSTM structure, and a multi-layer graph convolutional network (GCN) is layered on top of the LSTM. The primary function of the GCN is to model the application of syntactical dependency trees. The outputs of GCN are then passed through aspect specific masking and attention mechanisms. Finally, the results are scaled by softmax to produce a normalized probability distribution. The ASGCN demonstrated obvious advantage over existing baseline algorithms such as Support Vector Machine (SVM), LSTM, MemNet, Arithmetic Optimization Algorithm (AOA). An ablation study is also performed to assess the importance of each layer of the ASGCN, and GCN is found to be one of the most critical part of the ASGCN due to the fact that the removal of it caused a significant drop in accuracy. Further investigation is also conducted to dig into some of the details of the GCN, specifically on the number of layers of GCN and increased number of aspects in the data, and 2 layers seem to be optimal in this context and additional layers do take a toll on the overall accuracy. It should also be mentioned that ASGCN can still be extended to include domain knowledge or to determine multiple aspects, which is not yet explored at the time this paper was authored.

2.2. Aspect-gated graph convolutional networks (AGGCN)

AGGCN includes an aspect gate design that is able to better encode the aspect information of the input to produce better accuracy [2]. The AGGCN architecture deploys two pieces of advances comparing to the ASGCN. The first piece is that the LSTM model added an aspect gate, strengthening the encoding for the aspect extraction. The second piece is that the attention mechanism is added with a retrieval-based attention which will retrieve semantically relevant aspect words. The entire model is trained via cross-entropy and L2-regularization. Studies similar to the ASGCN were also done to investigate the optimal numbers of layers for GCN structure, and the result was that 2 layers seem to be optimal for the GCN and suspects more layers will tend to result in overfitting. The performance of AGGCN is a bit mixed when it comes to various datasets, and AGGCN outperforms Data Mining Template Library (DMTL) with Rest14 and AS-GCN with Rest16, but it did lose out to AS-GCN with Lap14 and DMTL with Rest15; the author attributes the shortcomings in the AGGCN's outcomes to the inherent properties of datasets [9].

An ablation study is also performed, and the results show a general drop in accuracy when removing components of the model, but the removal of aspect-gate and GCN does seem to improve the accuracy with Lap14 and Twitter which can be attributed to the aspect sensitivity and position information. The ablation study also revealed the components' impact on syntactical dependencies and sentiment dependency, and the case study that followed further confirmed the model's capabilities in these two areas.

2.3. Dependency graph enhanced dual-transformer network (DGEDT)

DGEDT is proposed to address two shortcomings of directly applying the dependency tree which is used by GCNs [3]. The first issue is that noisy information will be introduced, and the second issue is that the GCN is inherently inferior when working with a dependency tree in this context. The architecture of the network is composed of three critical parts: first, an aspect-based encoder with bidirectional LSTM; second, a dual-transformer structure consisting of a bidirectional GCN and BiAffine transformation process; third, an aspect-based attention module. The model's outcomes were really good, as the DGEDT outperforms all the other methods across all the datasets, and with

Bidirectional Encoder Representations from Transformers (BERT) added to the LSTM structure, the results were further improved. The ablation study also found that the removal of BiAffine will cause a drop in accuracy, demonstrating the critical nature of the BiAffine structure. The case studies also showed that DGEDT is capable of achieving the proper balance between dependency graph enhanced BiGCN and traditional Transformer according to different situations. It should also be acknowledged that there are still potentials to improve the DGEDT which relates to domain-specific knowledge, edge-aware neural networks, etc.

2.4. Transformer Based Multi-Grained Attention Network (T-MGAN)

T-MGAN is proposed to address the two disadvantages of previous methods that combine neural networks such as CNN and RNN with attention mechanisms[4]. The first issue is that previous methods often lose useful information when the aspects encountered are within a phrase instead of a word because they cannot fully extract all the features within a phrase. The second issue is that the previous methods deploy single pooling mechanisms which serves to further understand the contexts and the aspects, but the single pooling will also lose information in the process. Architecture is as follows: the first layer is a feature extraction layer which is composed of a transformer encoder, tree transformer encoder. The extracted info is then fed into a multi-attention layer which is an interplay of aspect and context. The final stage of the processing is going through an output layer which is simply a softmax function. The overall loss function used is cross-entropy. In the experiments of the T-MGAN against all the baseline algorithms, the T-MGAN outperforms all of them in most cases, as the transformer structure is able to capture the aspect features and contexts features. The effectiveness analysis and the case studies also reflect on the critical function that the transformer structure serves. However, it was acknowledged that the T-MGAN needed to be improved when working with more colloquial datasets.

3. Background of Aspect based Sentiment Analysis

3.1. Datasets

The methods examined in this paper utilize a range of datasets to evaluate the performance. Specifically, five datasets will be the focuses of this paper: SemEval 2014 task 4 (LAP 14), SemEval 2015 task 12 (Rest 14) , SemEval 2016 task 5 of category laptop (Rest 15), SemEval 2016 task 5 of category restaurant (Rest 16) and Twitter Dataset . All the methods except T-MAGN were tested on all five datasets, but T-MAGN was only tested on REST15, REST16 and Twitter datasets[5-8].

3.2. Metrics

There are two metrics used to evaluate the results of the models. The first metric is accuracy which directly calculates the portion of the correctly predicted samples. The formula for calculating accuracy is given in Equation 1. TP stands for true positives; TN stands for true negatives; FP stands for false positives; FN stands for false negatives. The second metric is called F1 score which is a more useful benchmark as it can deal with unbalanced datasets. It can often be the case that the model has a high accuracy rate but a low F1 score, because the dataset itself may have an overwhelming number of samples from one category. The formula for calculating the F1 score is given in Equation 2. The higher the accuracy and F1 score, the better the performance of the model.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$F1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{2}$$

4. Neural Networks for ABSA

Neural networks serve a critical part in the field of ABSA. RNNs and CNNs can be directly applied in ABSA, and Graph Convolutional Network (GCN) and variants of RNNs play important roles in the novel methods summarized in this paper. This section introduces the principles of a specific variant of RNN which is called Gated Recurrent Unit (GRU) and GCN.

4.1. GRU neural network

GRU and other variants of RNN like LSTM were introduced to address the vanishing gradient problem faced by vanilla RNN[12]. When training the RNN, the gradient that is used to update the parameters of RNN gets increasingly small with more layers, and this issue will prevent the RNN from any further training. GRU also helps to improve the RNN's ability to deal with long term dependencies since it included mechanisms to deal with long term memory. The internal architecture is simpler than that of the LSTM, which improves the efficiency when training the model.

4.2. Graph Convolutional Network (GCN)

The first step of constructing GCNs is to find the normalized Laplacian Matrix of the graph which is often characterized as multisource and heterogeneous. The equation for the Laplacian Matrix is given in Equation 3, where L is the desired matrix, D is the degree matrix, and A is the adjacency matrix.

$$L = I_n - D^{-\frac{1}{2}}AD^{-\frac{1}{2}} \tag{3}$$

In order to perform convolution, Laplacian matrix needs to be decomposed to feature values and feature vectors. The calculation is given in Equation 4, where x represents the scalar at each node, V represents the feature vectors, γ represents feature values.

$$g_\theta^* x = V g(\gamma) V^T * x \tag{4}$$

However, feature decomposition is often found to be intractable in real-world calculations. Thus, approximations are needed in this case to reduce the computational complexity. While there are a number of ways to perform approximations, this paper introduces Chebyshev polynomials which are given in Equation 5.

$$T_k(x) = 2xT_{k-1}(x) - T_{k-2}(x) \tag{5}$$

Using first order form of the Chebyshev polynomials to approximate $g(\gamma)$, the simplified form is given in Equation 6[15].

$$g_\theta^* x \approx \theta(I_N + D^{-\frac{1}{2}}AD^{-\frac{1}{2}})^* x \tag{6}$$

Accordingly, for a two-layer GCN, each layer of propagation can be demonstrated in equation 7,8,9&10, where h^0 is the initial input to the first layer, and h^1 stands for the output of the first GCN layer, and h^2 is the output of the second layer, and σ is an activation function such as Rectified Linear Activation Unit (ReLU) or sigmoid, and Ws are the weights of each layer.

$$A^* = A + I_N \tag{7}$$

$$A^\wedge = D^{-\frac{1}{2}}A^*D^{-\frac{1}{2}} \tag{8}$$

$$h^1 = \sigma(A^\wedge h^0 W^1) \tag{9}$$

$$h^2 = A^\wedge h^1 W^2 \tag{10}$$

The outputs of the last GCN layer usually have to go through a processing function like softmax which depends on the desired representation of the output.

5. Comparison between different models

There are a number of results obtained across all the models. This paper gathered all the best performing results from each paper, and the gathered results are displayed in the Table 1. Best performing models generally either outperform baseline models or have shown outstanding results in specific datasets. While all the novel models generally outperforms baselines, there are differences when it comes to variances of these novel models, and some baseline models also demonstrated surprisingly good outcomes. More specifically, 1) MGAN is added here as it scores higher in the twitter dataset comparing to T-MGAN, and the colloquial contexts is of high importance of the studies. It also should be mentioned that the T-MGAN models did not have experimental data on the Lap14 and Rest14 dataset, and future work should be done in this regard to get a better sense on T-MGAN comparing to other competing models, especially DGEDT-BERT[12-13]. 2) DGEDT-BERT is variation of DGEDT with an added BERT layer, and these two models show overwhelmingly good results across all data sets. 3) ASGCN has two variants: ASGCN-DT and ASGCN-DG. There are still performance differences within these variants. While ASGCN-DG outperforms ASGCN-DT across LAP14 and REST15, the ASGCN-DT does win when it comes to REST14. 4) DMTL is added on here for a couple of reasons: firstly, it outperforms AGGCN and ASGCN in the REST15 dataset; secondly, DMTL scored similar levels of results compared to the AGGCN in the LAP14 dataset and outperformed ASGCN in the REST14 dataset; finally, the experiments with DMTL on Twitter dataset was not done, and it remains to be seen whether DMTL can score a surprisingly high performance in more colloquial contexts. 5)The TNet-LF which is a baseline model has higher accuracy than both of the variants on the twitter dataset and REST16. It should be noted that even though the datasets used in these experiments are the same, there are still differences in performances in the baseline models such as LSTM. Such differences can be attributed to implementation and training parameters. The best two rows of each column are highlighted in Table 1.

Table 1. Results of best performing models

	Lap14		Rest14		Rest15		Rest16		Twitter	
	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1
T-MGAN	/	/	/	/	76.38	**73.02**	82.06	72.65	71.23	70.63
MGAN	/	/	/	/	75.39	**72.47**	81.25	71.94	72.54	70.81
DGEDT	**76.8**	**72.3**	83.9	**75.1**	82.1	65.9	**90.8**	**73.8**	**74.8**	**73.4**
DGEDT-BERT	**79.8**	**75.6**	**86.3**	**80**	**84**	71	**91.9**	**79**	**77.9**	**75.4**
AGGCN	73.53	68.99	**84.37**	73.82	80.27	64.51	90.53	73.92	73.64	72.2
DMTL	73.35	68.93	82.23	72.49	**82.72**	66.94	88.13	71.48	/	/
ASGCN-DG	75.55	71.05	80.77	72.02	79.89	61.89	88.99	67.48	72.15	70.4
ASGCN-DT	74.14	69.24	80.86	72.19	79.34	60.78	88.69	66.64	71.53	69.68
TNet-LF	74.61	70.14	80.42	71.03	78.47	59.47	89.07	70.43	72.98	71.43

The DGEDT-BERT model ranks the highest amongst all the other algorithms in every dataset except REST15. While DGEDT-BERT does have a higher accuracy rate, the F-1 scores of T-MGAN and MGAN are dramatically higher. The data description of REST15 is shown in the Table 2.

Table 2. Data Description of REST15

REST15	positive	negative	neutral
Train	978	307	36
Test	326	182	34

Since the positive samples are significantly more than the negative samples, F1 scores seem to be a better benchmark since it is more fitted for unbalanced datasets. It is thus reasonable to conclude that T-MGAN and MGAN do have an edge against the DGEDT-BERT in the REST15 dataset [14].

Future research should be done to study why this is the case. Since REST15 and REST16 were first introduced as the same category of data with different focuses on restaurants and laptops, one may expect that same model should have similar performances on them. However, the results here revealed a different phenomenon. The model with the second highest accuracy, DMTL, showed a drastically lower F1 score, which can be explained partly by the imbalanced nature of the dataset. The DGEDT-BERT model, on the other hand, still demonstrated comparable levels of F1 compared to T-MGAN, but it outperforms disproportionally better than T-MAGN in the REST16 which is supposed to be inherently the same as REST15. It should also be mentioned that DGEDT is the second-best performing model after DGEDT-BERT, which illustrated the superiority of DGEDT, and adding BERT not only confirmed the effectiveness of BERT in the area of ABSA but also opened up new possibilities of integrating BERT into other models.

6. Challenges

6.1. Establishing standard benchmarks for baseline models

It is often the case that a novel method would be compared against several baseline models like LSTM and SVM. While in principle the same model trained with the same dataset should yield the same results, the results of baseline models reviewed in this paper did show some discrepancies in consistency across various research. Even though comparing different models on the same dataset with this discrepancy still reveals important implications, removing such discrepancy is crucial to solidify the credibility of the novel models [15].

One solution to this issue is to conduct research to optimally train various baseline models on specific datasets, and the procedures of training and testing should be documented in way that is easy to reproduce, and this will establish a benchmark performance for baseline models. This solution will not only solve the issue of inconsistency but also simplify the tasks for future researchers because they can simply borrow the standard results from the benchmark.

6.2. Better categorization of the datasets

Datasets can be inherently different due to the way they were created, and models perform differently with different datasets. For example, T-MAGN loses out in the Twitter dataset while wins on others, and the explanation for such phenomenon was that the twitter dataset is more colloquial.

Several problems arise from this explanation. Firstly, it was unclear when a dataset should be categorized as colloquial, or if being colloquial is a clear classification or possesses various degrees. Secondly, it was unclear if being colloquial is the only factor that influences the performance of models, and there might be others that requires better categorizations of datasets to explore. Systematic research on this challenge will offer a much better sense to future research about which model works best in what context.

6.3. Identifying new useful components

Some components, when integrated into existing architectures of the models, can improve the performances of the model in general. BERT, for instance, boosted the performances of DGEDT overall. More importantly, BERT can be coupled with other models for ABSA as well which have

not been thoroughly investigated. Identifying such components and introducing these components into the existing model can not only increase the performance but also opens up new research possibilities.

7. Conclusions

The novel approaches that have been proposed does dramatically outperforms traditional RNN and CNN as well as baseline models like SVM. While novel methods do show a clear advantage, baseline models such DMTL can be better in some scenarios as well. This paper introduced four models: ASGCN AGGCN, DGEDT and T-MGAN, and DGEDT when boosted by BERT seems to be the best model amongst all the models that this paper has examined so far. It should be noted that the potentials of T-MAGN need to be fully explored in other datasets, and REST15 and REST 16 datasets need to be further evaluated to address the anomalies identified in this paper.

References

[1] Zhang, C., Li, Q., & Song, D. (2019, November). Aspect-based Sentiment Classification with Aspect-specific Graph Convolutional Networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP) (pp. 4568-4578).

[2] Lu, Q., Zhu, Z., Zhang, G., Kang, S., & Liu, P. (2021). Aspect-gated graph convolutional networks for aspect-based sentiment analysis. Applied Intelligence, 1-12.

[3] Tang, H., Ji, D., Li, C., & Zhou, Q. (2020, July). Dependency graph enhanced dual-transformer structure for aspect-based sentiment classification. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 6578-6588).

[4] Sun, J., Han, P., Cheng, Z., Wu, E., & Wang, W. (2020). Transformer Based Multi-Grained Attention Network for Aspect-Based Sentiment Analysis. IEEE Access, 8, 211152-211163.

[5] Dong, L., Wei, F., Tan, C., Tang, D., Zhou, M., & Xu, K. (2014, June). Adaptive recursive neural network for target-dependent twitter sentiment classification. In Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 2: Short papers) (pp. 49-54).

[6] JPHPIASM Maria Pontiki, D. G. (2014). Semeval-2014 task 4: Aspect based sentiment analysis. In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), RecSys (Vol. 14).

[7] Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. Semeval-2015 task 12: Aspect based sentiment anal- ysis. In Proceedings of the 9th

[8] Maria Pontiki, Dimitris Galanis, Haris Papageor- giou, Ion Androutsopoulos, Suresh Manandhar, AL-Smadi Mohammad, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphe´e De Clercq, et al. 2016. Semeval-2016 task 5: Aspect based sentiment anal- ysis. In Proceedings of the 10th international work-shop on semantic evaluation (SemEval-2016), pages 19–30.

[9] Al Hasan, M., Chaoji, V., Salem, S., Parimi, N., & Zaki, M. J. (2005). DMTL: A generic data mining template library. Library-Centric Software Design (LCSD'05), 53.

[10] Li, X., Bing, L., Lam, W., & Shi, B. Transformation Networks for Target-Oriented Sentiment Classification.

[11] Pascual, F. (2019, March 8). Guide to Aspect-Based Sentiment Analysis. MonkeyLearn Blog. https://monkeylearn.com/blog/aspect-based-sentiment-analysis/

[12] Cho, K., van Merriënboer, B., Gulcehre, C., Schwenk, F. B. H., & Bengio, Y. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation.

[13] Xue, W., & Li, T. (2018, July). Aspect Based Sentiment Analysis with Gated Convolutional Networks. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 2514-2523).

[14] Bahdanau, D., Cho, K., & Bengio, Y. (2017). Neural Machine Translation by Jointly Learning to Align and Translate.

[15] Kipf, T. N., & Welling, M. SEMI-SUPERVISED CLASSIFICATION WITH GRAPH CONVOLUTIONAL NETWORKS.

[16] Do, H. H., Prasad, P., Maag, A., & Alsadoon, A. (2019). Deep Learning for Aspect-Based Sentiment Analysis: A Comparative Review. Expert Systems with Applications, 118, 272–299. https://doi.org/10.1016/j.eswa.2018.10.003