Data Augmentation Using BERT-Based Models for Aspect-Based Sentiment Analysis

Bron Hollander, Flavius Frasincar[0000−0002−8031−758X](✉), and Finn van der Knaap

Erasmus University Rotterdam, Burgemeester Oudlaan 50, 3062 PA Rotterdam, the Netherlands
511709bh@student.eur.nl, frasincar@ese.eur.nl, 573834fk@student.eur.nl

Abstract. In today's digital world, there is an overwhelming amount of opinionated data on the Web. However, effectively analyzing all available data proves to be a resource-intensive endeavor, requiring substantial time and financial investments to curate high-quality training datasets. To mitigate such problems, this paper compares data augmentation models for aspect-based sentiment analysis. Specifically, we analyze the effect of several BERT-based data augmentation methods on the performance of the state-of-the-art HAABSA++ model. We consider the following data augmentation models: EDA-adjusted (baseline), BERT, Conditional-BERT, $\text{BERT}_{\text{prepend}}$, and $\text{BERT}_{\text{expand}}$. Our findings show that incorporating data augmentation techniques can significantly improve the out-of-sample accuracy of the HAABSA++ model. Specifically, our results highlight the effectiveness of $\text{BERT}_{\text{prepend}}$ and $\text{BERT}_{\text{expand}}$, increasing the test accuracy from 78.56% to 79.23% and from 82.62% to 84.47% for the SemEval 2015 and SemEval 2016 datasets, respectively.

Keywords: Aspect-based sentiment classification · Data augmentation · Neural network · Pre-trained language model

1 Introduction

The modern era of the Web has made it effortless for people to share information, allowing consumers to express their opinions about various products and services more easily than ever. The abundance of user-generated data presents an opportunity for consumers and businesses. For instance, businesses could use newly created reviews to confirm their marketing strategy at several levels [6], whereas consumers could use it to help them make more informed decisions [20]. However, effectively using the available data requires a deep understanding of the contents and sentiment present in the review. As such, Aspect-Based Sentiment Analysis (ABSA), which entails extracting the sentiment with respect to an aspect in a review, can be valuable. According to the survey of [1], ABSA encompasses three primary approaches: a knowledge-based approach, a machine-learning approach, and a hybrid approach. [1] also demonstrates the potential

of hybrid models to effectively predict sentiment. Nonetheless, a common issue of these models is the lack of available labeled data for training purposes.

To address the scarcity of labeled data, previous literature has proposed several data augmentation techniques [14]. [8] shows the effectiveness of Easy Data Augmentation (EDA) in improving sentiment predictions of the Hybrid Approach for Aspect-Based Sentiment Analysis (HAABSA) model [16]. Nevertheless, EDA has its limitations, such as potential changes in sentiment or sentence incoherence after augmentation. Recent studies use neural networks for data augmentation, in particular pre-trained transformer models, to enhance sentiment preservation and contextual awareness during augmentation [5].

The impressive performance of Bidirectional Encoder Representations from Transformers (BERT)-based models in Natural Language Processing (NLP) tasks suggests that such models may be well suited for data augmentation. In this paper, we aim to investigate the impact of BERT-based data augmentation techniques on the performance of the HAABSA++ model, a state-of-the-art hybrid method for ABSA proposed in [15].

The contribution of this paper to existing literature is as follows. In contrast to previous approaches, such as EDA for HAABSA [8], we extend the comparison of data augmentation techniques for HAABSA++ to Pre-trained Language Models (PLMs) [5], namely BERT [2], Conditional-BERT (C-BERT) [18], $\text{BERT}_{\text{prepend}}$, and $\text{BERT}_{\text{expand}}$ [5], therefore providing a homogeneous comparison between all aforementioned data augmentation models. This paper focuses on BERT instead of other language models, as [18] shows the superior effectiveness of a bidirectional language model over a unidirectional language model.

The Python source code and data (SemEval 2015 [12] and 2016 [13] restaurant review datasets) used in this study are available at https://github.com/BronHol/HAABSA_PLUS_PLUS_DA. Figure 1 illustrates an example review represented in the XML format.

```
▼<sentence id="1032695:1">
   <text>Everything is always cooked to perfection, the service is excellent, the decor cool and understated.</text>
  ▼<Opinions>
     <Opinion target="NULL" category="FOOD#QUALITY" polarity="positive" from="0" to="0"/>
     <Opinion target="service" category="SERVICE#GENERAL" polarity="positive" from="47" to="54"/>
     <Opinion target="decor" category="AMBIENCE#GENERAL" polarity="positive" from="73" to="78"/>
   </Opinions>
 </sentence>
```

Fig. 1. Example review in the XML format

The rest of the paper is structured as follows. Section 2 gives an overview of the HAABSA++ model and the considered data augmentation techniques followed by, in Sect. 3, a discussion of the obtained results. Last, Sect. 4 provides our conclusion and suggestions for future research.

2 Methodology

In this section, we present the HAABSA++ model and the considered methods for data augmentation. First, we discuss the model used for ABSA, HAABSA++, in Subsect. 2.1, which is a two-step approach that combines a domain sentiment ontology and a neural network. Second, the considered data augmentation techniques are presented in Subsect. 2.2.

2.1 HAABSA++

HAABSA++ [15] is a hybrid approach that extends the HAABSA model [16] with contextualized word embeddings and hierarchical attention. The first step of HAABSA++ is to classify the sentiment using a domain sentiment ontology. For the inconclusive cases, HAABSA++ uses the LCR-Rot-hop++ model.

Domain Sentiment Ontology. The used domain sentiment ontology in the HAABSA++ model [15] predicts sentiment by leveraging predefined classes, class relations, and axioms. It is important to note that the ontology reasoner does not incorporate any classes or relations specifically addressing neutral sentiment. As such, the rule-based methodology is limited to detecting positive and negative emotions. Therefore, the ontology may not be reliable in three situations: (1) neutral sentiment, which is intentionally excluded from the reasoner, (2) conflicting sentiment, where both positive and negative sentiments are predicted for a target, and (3) no hits, resulting from limited coverage of the ontology. In such cases, the LCR-Rot-hop++ model serves as a backup method.

LCR-Rot-hop++. The LCR-Rot-hop++ model [15] is an extension of the LCR-Rot model [21]. [16] first improves the model by repeating the rotatory attention mechanism which helps to properly weight the relevant sentiment words, resulting in LCR-Rot-hop. [15] then replaces the context-independent GloVe embeddings [11] with context-dependent BERT embeddings, and proposes a hierarchical attention structure to enhance the model's flexibility, resulting in the LCR-Rot-hop++ model.

2.2 Data Augmentation

Data augmentation involves manipulating existing data to expand the size of a dataset artificially, thereby generating additional data points with modified variations. The expansion of training data through data augmentation is crucial for improving the sentiment predictions of HAABSA++, or more specifically LCR-Rot-hop++, due to the limited available training data. In this section, we discuss the various data augmentation techniques considered in this paper.

Easy Data Augmentation. The EDA technique, proposed in [17], is a straightforward and effective approach for data augmentation in NLP tasks. It encompasses four operations: synonym replacement, random insertion, random swap, and random deletion. [8] extends the EDA method specifically for ABSA tasks, adding word sense disambiguation, which tackles the challenge of selecting the correct word meaning and function within a sentence. The POS tag of each word is determined, and the Lesk algorithm is used to identify the most suitable word meaning based on contextual information. The simplified Lesk algorithm, implemented using the WordNet library, is used for both synonym replacement and random insertion. Moreover, the proposed method introduces target swaps across sentences, enabling the swapping of target words within the same category to provide diverse contexts. The EDA-adjusted model, combining these three methods, serves as a baseline model in this paper.

BERT. A more advanced approach for data augmentation in ABSA involves using PLMs. BERT [2] is advantageous because it captures both the left and right context simultaneously and therefore considers the context of the target word [3], making it extremely useful for ABSA tasks. BERT is trained using MLM and Next Sentence Prediction (NSP). In MLM, certain words in a sentence are masked, and BERT tries to predict the masked words. NSP involves providing BERT with two sentences (A and B) and asking the model to predict whether sentence B follows sentence A. The input embeddings of BERT consist of token embeddings, segment embeddings, and position embeddings [2].

To augment the data, we use the MLM task of BERT [2]. During MLM, multiple candidate words are generated as potential replacements for masked tokens. This approach generates new sentences that convey similar meaning to the original input. Following the standard BERT approach, we mask each word in a sentence with a probability of 15% [5]. Subsequently, we select the substitute word with the highest probability for each masked word, excluding the original word. By applying this process to every sentence in the original training dataset, we generate a new sentence for each sentence in the dataset.

C-BERT. A downside of using BERT for data augmentation is that the original sentiment label of a sentence is not taken into account, which may result in the loss of the original sentiment information when replacing the masked words. To address this issue, [18] proposes the C-BERT model. In the C-BERT model, the sentence is augmented conditional on the label of the sentence itself. To incorporate label information during the MLM process, the segment embeddings in BERT are replaced with label embeddings, and the model is trained on labeled datasets. Once C-BERT is trained and equipped with knowledge of both sentiment and context, it can be used to augment data similarly to the original BERT model.

$\text{BERT}_{\text{prepend}}$. By replacing the segment embedding with label embeddings, C-BERT becomes less suitable for diverse tasks because of its inherent speci-

ficity. [5] introduces $\text{BERT}_{\text{prepend}}$ as an extension of the original BERT model to condition data augmentations on the label without sacrificing generality. In $\text{BERT}_{\text{prepend}}$, the label of each sequence is prepended to the sequence itself, without including the label in the model's vocabulary. By considering the label of a sequence, $\text{BERT}_{\text{prepend}}$ facilitates data augmentations that are label-aware. The label of the sequence remains fixed, ensuring that it is not masked during augmentation. After applying $\text{BERT}_{\text{prepend}}$, the labels are removed, and the augmented data is incorporated into the training data of HAABSA++.

$\text{BERT}_{\text{expand}}$. $\text{BERT}_{\text{expand}}$ follows a similar approach to $\text{BERT}_{\text{prepend}}$, where the label of each sequence is prepended to the sequence itself. However, a notable difference is that $\text{BERT}_{\text{expand}}$ includes the label in the model's vocabulary, unlike $\text{BERT}_{\text{prepend}}$. In $\text{BERT}_{\text{expand}}$, the label is treated as a single token, whereas $\text{BERT}_{\text{prepend}}$ may split it into multiple subwords depending on the used word tokenizer [4].

Fine-tuning BERT-based Models. We fine-tune the hyperparameters of our models for the MLM task and use the complete SemEval training set. For $\text{BERT}_{\text{prepend}}$ and $\text{BERT}_{\text{expand}}$, we prepare the dataset by prepending the sentiment label of each sequence to the sequence. 80% of the training dataset is used for fine-tuning and the other 20% is used for validation of hyperparameter configurations. We run the fine-tuning process for 10 epochs. Following [5], we use the default masking parameters.

3 Results

The training and test accuracies of the considered models are presented in Table 1, including the number of data augmentations added to the training data for each model. The training accuracy score pertains to the in-sample accuracy, while the testing accuracy corresponds to the out-of-sample accuracy. The inclusion of training scores primarily serves to gauge potential model overfitting, while the evaluation of model performance relies primarily on testing accuracies.

For the SemEval 2015 dataset, the HAABSA++ model without any data augmentation achieves a test accuracy of 78.56%. Comparatively, the EDA-adjusted model achieves the highest test accuracy of 82.41%, which is an improvement of 3.85 percentage points. The EDA-adjusted model uses 3834 data augmentations, whereas all the BERT-based models only use 1278 data augmentations (EDA-adjusted has three augmentation equations while BERT-based models only have one). The best-performing BERT-based models are $\text{BERT}_{\text{prepend}}$ and $\text{BERT}_{\text{expand}}$, with both a test accuracy of 79.23%. Contrary to the other data augmentation models, we observe that C-BERT does not improve the performance of the HAABSA++ model.

For the SemEval 2016 dataset, we observe that the plain HAABSA++ model achieves a test accuracy of 82.62%. However, for this dataset, the EDA-adjusted

Table 1. The training and test accuracies of HAABSA++ and the considered data augmentation models

	SemEval 2015			**SemEval 2016**		
	Train acc.	Test acc.	#aug.	Train acc.	Test acc.	#aug.
HAABSA++	90.86%	78.56%	0	89.96%	82.62%	0
EDA-adjusted	90.70%	**82.41%**	3834	89.75%	81.85%	5640
BERT	91.02%	79.06%	1278	**91.30%**	82.77%	1880
C-BERT	**91.12%**	75.71%	1278	90.91%	82.00%	1880
$\text{BERT}_{\text{prepend}}$	91.04%	79.23%	1278	89.29%	**84.47%**	1880
$\text{BERT}_{\text{expand}}$	91.04%	79.23%	1278	89.29%	**84.47%**	1880

model does not increase the test accuracy compared to HAABSA++, obtaining a test accuracy of 81.85% despite using 5640 augmentations. Similarly, the C-BERT model does not improve the performance of HAABSA++. The poor out-of-sample performance of C-BERT in both datasets could be attributed to the replacement of segment embeddings with label embeddings, thereby forgetting the order of sentences and the semantics between these sentences. On the other hand, the BERT model achieves a test accuracy of 82.77% with 1880 augmentations, whilst both $\text{BERT}_{\text{prepend}}$ and $\text{BERT}_{\text{expand}}$ outperform all other models with a test accuracy of 84.47%. So, $\text{BERT}_{\text{prepend}}$ and $\text{BERT}_{\text{expand}}$ obtain an improvement of 1.85 percentage points.

EDA-adjusted, which is a lexicon-based method using grammatical rules and linguistics, works better for smaller datasets. On the other hand, machine learning approaches, such as the considered BERT-based models, thrive with larger datasets. As a result, EDA-adjusted achieves the highest out-of-sample accuracy for the SemEval 2015 dataset but performs modestly on the SemEval 2016 dataset.

An intriguing observation is that $\text{BERT}_{\text{prepend}}$ and $\text{BERT}_{\text{expand}}$ yield identical results. This can be attributed to the WordPiece tokenizer [19] that is used in BERT. Due to the tokenizer's behavior, sentiment labels (positive, neutral, and negative) remain intact without being split into multiple tokens. Consequently, incorporating the sentiment labels into the tokenizer's vocabulary does not alter the tokenization process for sentiment labels. As a result, the fine-tuning process treats the prepended labels in the same manner, resulting in indistinguishable data augmentations.

4 Conclusion

In this work, we extended the state-of-the-art HAABSA++ model proposed in [15] by incorporating various data augmentation techniques, including EDA-adjusted [8] and BERT-based models [5,18]. The main objective of data augmentation is to enhance the out-of-sample accuracy by training the neural network of HAABSA++ more effectively on a larger training dataset. Our findings

revealed that the performance of HAABSA++ can indeed be improved upon through the use of data augmentation methods, although the effectiveness of each data augmentation model varies depending on the used dataset. Specifically, for the smaller SemEval 2015 dataset, the lexicon-based EDA-adjusted method achieved the largest improvement, with an increase of 3.85 percentage points over the baseline. On the other hand, for the larger SemEval 2016 dataset, the $\text{BERT}_{\text{prepend}}$ and $\text{BERT}_{\text{expand}}$ methods performed best, with an increase of 1.85 percentage points. Overall, based on the performance for both datasets, $\text{BERT}_{\text{prepend}}$ and $\text{BERT}_{\text{expand}}$ emerge as the most effective data augmentation methods for the HAABSA++ model.

For future research, it could be interesting to examine the impact of selectively masking words in the MLM process. Now, we observe that words without semantic information are substituted by the MLM task. [10] shows that masking sentimental words or adjectives and adverbs can lead to improvements in performance. In addition, the inclusion of additional data augmentation models, such as BART or RoBERTa, could be interesting, as these models have demonstrated excellent performance in a variety of tasks [7,9].

References

1. Brauwers, G., Frasincar, F.: A survey on aspect-based sentiment classification. ACM Computing Surveys **55**(4), 65:1–65:37 (2023)
2. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019). pp. 4171–4186. ACL (2019)
3. Hoang, M., Bihorac, O.A., Rouces, J.: Aspect-based sentiment analysis using BERT. In: 22nd Nordic Conference on Computational Linguistics (NoDaLiDa 2019). pp. 187–196. Linköping University Electronic Press (2019)
4. Kudo, T., Richardson, J.: SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In: 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018). pp. 66–71. ACL (2018)
5. Kumar, V., Choudhary, A., Cho, E.: Data augmentation using pre-trained transformer models. arXiv preprint arXiv:2003.02245 (2020)
6. Lee, T.Y., Bradlow, E.T.: Automated marketing research using online customer reviews. Journal of Marketing Research **48**(5), 881–894 (2011)
7. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020). pp. 7871–7880. ACL (2020)
8. Liesting, T., Frasincar, F., Trusca, M.M.: Data augmentation in a hybrid approach for aspect-based sentiment analysis. In: 36th ACM/SIGAPP Symposium on Applied Computing (SAC 2021). pp. 828–835. ACM (2021)
9. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692 (2019)

10. Pantelidou, K., Chatzakou, D., Tsikrika, T., Vrochidis, S., Kompatsiaris, I.: Selective word substitution for contextualized data augmentation. In: 27th International Conference on Applications of Natural Language to Information Systems (NLDB 2022). LNCS, vol. 13286, pp. 508–516. Springer (2022)
11. Pennington, J., Socher, R., Manning, C.D.: GloVe: Global vectors for word representation. In: 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014). pp. 1532–1543. ACL (2014)
12. Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., Androutsopoulos, I.: SemEval-2015 task 12: Aspect based sentiment analysis. In: 9th International Workshop on Semantic Evaluation (SemEval 2015). pp. 486–495. ACL (2015)
13. Pontiki, M., Galanis, D., Papageorgiou, H., Manandhar, S., Androutsopoulos, I., Al-Smadi, M., Al-Ayyoub, M., Zhao, Y., Qin, B., Clercq, O.D., Hoste, V., Apidianaki, M., Tannier, X., Loukachevitch, N.V., Kotelnikov, E.V., Bel, N., Zafra, S.M.J., Eryigit, G.: SemEval-2016 task 5: Aspect based sentiment analysis. In: 10th International Workshop on Semantic Evaluation (SemEval 2016). pp. 19–30. ACL (2016)
14. Shani, C., Zarecki, J., Shahaf, D.: The lean data scientist: Recent advances toward overcoming the data bottleneck. Communications of the ACM **66**(2), 92–102 (2023)
15. Trusca, M.M., Wassenberg, D., Frasincar, F., Dekker, R.: A hybrid approach for aspect-based sentiment analysis using deep contextual word embeddings and hierarchical attention. In: 20th International Conference of Web Engineering (ICWE 2020). LNCS, vol. 12128, pp. 365–380. Springer (2020)
16. Wallaart, O., Frasincar, F.: A hybrid approach for aspect-based sentiment analysis using a lexicalized domain ontology and attentional neural models. In: 16th Extended Semantic Web Conference (ESWC 2019). LNCS, vol. 11503, pp. 363–378. Springer (2019)
17. Wei, J.W., Zou, K.: EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In: 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP 2019). pp. 6381–6387. ACL (2019)
18. Wu, X., Lv, S., Zang, L., Han, J., Hu, S.: Conditional BERT contextual augmentation. In: 19th International Conference on Computational Science (ICCS 2019). LNCS, vol. 11539, pp. 84–95. Springer (2019)
19. Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al.: Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144 (2016)
20. Zhang, Y., Du, J., Ma, X., Wen, H., Fortino, G.: Aspect-based sentiment analysis for user reviews. Cognitive Computation **13**(5), 1114–1127 (2021)
21. Zheng, S., Xia, R.: Left-center-right separated neural network for aspect-based sentiment analysis with rotatory attention. arXiv preprint arXiv:1802.00892 (2018)