Text Data Augmentation Method Using Filtering Indicators based on Multiple Perspectives

Haruto Uda[1], **Kazuyuki Matsumoto**[1], **Minoru Yoshida**[1]
[1]Division of Science and Technology,
Graduate School of Sciences and Technology for Innovation,
Tokushima University, Japan
c612335004@tokushima-u.ac.jp, {matumoto; mino}@is.tokushima-u.ac.jp

Abstract

The widespread use of social networking services (SNS) has made it possible to collect a wide variety of text data on a large scale. Text data posted on SNS contain many broken expressions, especially abbreviations and colloquial expressions. In order to utilize such data as a resource for natural language processing, annotation of the data, assignment of class labels, etc. become issues. In general, because manual annotation is costly, artificial data augmentation and semi-automation of label assignment are often used as a countermeasure against data shortages. In this study, we propose a method for efficiently preparing large-scale, high-quality labeled text data for machine learning by applying evaluation indicators from multiple perspectives to the data generated by data augmentation methods. The goal is to improve the prediction accuracy of the model by adding the augmented data to the training data. Specifically, the proposed method sets thresholds for the semantic similarity based on the vector of BERT between the original text and the augmented text, the degree of change by BLEU, and the change in attention by Attention, respectively, and deletes data that do not satisfy the threshold conditions. Since the number of augmented data also affects learning accuracy, the number of data is also addressed by adding it to the evaluation indicators. Evaluation experiments on emotion-labeled datasets show that the proposed method achieves higher Accuracy than the method that simply augments the data using Easy Data Augmentation.

1 Introduction

In recent years, it has become easy to obtain vast and diverse text data on the World Wide Web (Web). However, there are several problems with text data on the Web. For example, text data posted on social networking services (SNS) tend to be short sentences with abbreviations, slang, colloquialisms, and colloquial expressions, reducing the number of words needed for people to grasp the meaning of a sentence. This makes consistent labeling difficult in the creation of training data for natural language processing tasks. In addition, manually preparing large, high-quality, labeled text data for machine learning is generally expensive. Data augmentation methods exist as an efficient way to prepare training data without human intervention. Data augmentation is the automatic generation of different data that are similar by performing various processes on the data so as not to spoil its essence. This can be expected to improve the prediction accuracy of the model.

In order to improve the learning accuracy of sentiment classification of text data, this research aims to increase the number and quality of training data by applying evaluation indicators from multiple perspectives to the data generated by the data augmentation method. When data augmentation is easily applied to text, it may cause a significant change in the meaning of the text, which may result in a loss of accuracy. For example, in image data augmentation, operations such as blurring, inversion, and color change can generate a large amount of effective training data. However, with text, a single missing word or a change in the order of words can drastically change the meaning of a sentence. Therefore, the augmentation process is likely to generate meaningless text or text that belongs to different classes, which may cause accuracy loss. As a data augmentation method, Easy Data Augmentation (EDA) by Wei and Zou (2019). is used to deal with data imbalances and shortages by generating multiple texts from a single text. In addition, in order to avoid inappropriate text for training data, which causes the aforementioned accuracy loss, we investigate how to suppress the loss of learning accuracy by applying evaluation indicators to the text generated by the data augmentation method.

2 Related Work

Wei and Zou (2019) proposed Easy Data Augmentation (EDA) as a method for augmenting simple text data in English. The main data augmentation operations on text data with EDA are synonym replacement, synonym insertion, word movement, and word deletion. Classification experiments using deep learning with SST-2(Socher et al., 2013), CR(Hu and Liu, 2004), SUBJ(Pang and Lee, 2004), TREC(Li and Roth, 2002), and PC(A. and Miller, 1995) datasets were conducted and showed great effectiveness when the number of original datasets was small. In this research, EDA is applied to the Japanese language and data augmentation is performed.

Okimura et al. (2022) used 12 different data augmentation methods with pre-trained models. MRPC(Dolan and Brockett, 2005), SICK(Marelli et al., 2014), and SST-2(Socher et al., 2013) were used for the dataset. The performance improvement was confirmed when using a dataset of several hundred examples, suggesting the effectiveness of data augmentation when training with a pre-trained model. Cosine similarity and BLEU were used to evaluate the sentences generated by data augmentation, and their impact on learning was analyzed. In this research, we evaluate the text generated by data augmentation to find the optimal threshold of evaluation values for the training data.

Yamada et al. (2022) proposed a method for adaptively selecting a data augmentation method utilizing Transformer(Vaswani et al., 2017) for image data. The Transformer can learn through its internal Self-Attention mechanism to obtain appropriate weights for its inputs. By using this Attention, the appropriate data was analyzed from the augmented data. In this research, Attention is used as a evaluation indicator of data augmentation for the text data.

Uda et al. (2023) performed data augmentation on Japanese text data, and selected the augmented data according to the evaluation indicators of the augmented data using cosine similarity and BLEU. As a result, the classification accuracy of the model was improved by manipulating the threshold of the evaluation indicators. In this research, we aim to improve the quality of augmented data by adding Attention, a new evaluation indicator, to cosine similarity and BLEU.

3 Method

In this section, we describe the dataset used and the proposed method. The Figure1 shows the flow of this research.

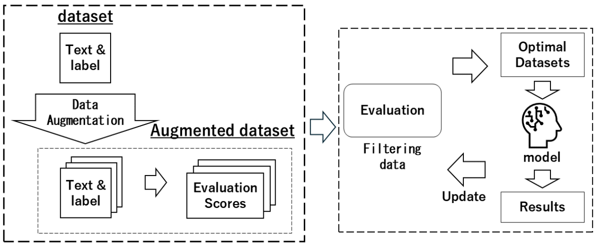

Figure 1: Flow of this research. The left side shows the flow of data augmentation, and the right side shows the flow of learning and searching for the optimal value of the evaluation indicators.

3.1 Dataset

For the dataset, we use WRIME corpus created by Kajiwara et al. (2021) as a reliable source of assigned labels. This corpus consists of past posted texts on SNS to which emotional intensity has been assigned by the posters themselves and by readers, both subjectively and objectively. The labels used in the experiment are the five emotional polarities of the WRIME corpus: strong positive, strong negative, positive, negative, and neutral, and three emotional polarities: positive, negative, and neutral.

3.2 Data Augmentation

The Figure2 briefly illustrates the data augmentation process.The data augmentation of the text included synonym replacement (SR), synonym insertion (SI), word swap (WS), and word deletion (WD). In EDA, changes in sentence meaning were suppressed by using stop words in word selection. In this research, data augmentation is performed for all words in order to suppress changes in sentence meaning and increase text expandability through evaluation indicators. In the process, MeCab(Kudo et al., 2004) was used to separate Japanese words into phrases. The Japanese WordNet(Yamada et al., 2010) developed by the National Institute of Information and Communications Technology (NICT) is used for synonym selection. The Japanese WordNet is a Japanese semantic dictionary that has a set of synonym relations for words, and we randomly selects words from the set of synonyms.

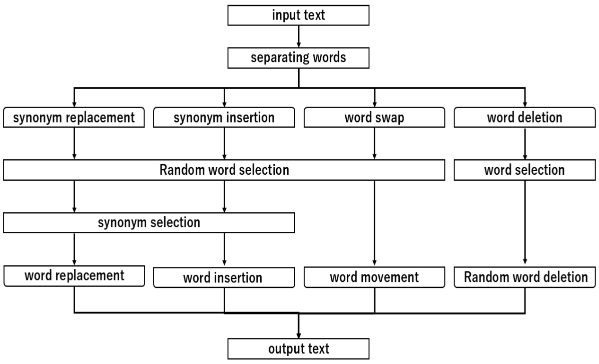

Figure 2: Process steps of data expansion. For a single text input, text is generated using four augmentation methods. The generated text is output as a set of augmented data.

3.3 Evaluation

Thresholds are set for the semantic similarity based on the vector of BERT between the original text and the augmented text, the degree of change by BLEU, and the change in attention by Attention, respectively, and data that do not satisfy the threshold conditions are removed. In this way, similar texts and texts that are not appropriate for the label of the data are filtered out to avoid deterioration of the quality of the training data. The following are the evaluation indicators used in the experiments.

- Semantic Similarity (SS) :Cosine similarity between CLS vectors from learned BERT of original and augmented text
- Degree of Text Change (DTC) :BLEU score between original and augmented text
- Word Attention (WA) :Sum of the difference in attention between corresponding words in the original and augmented text

3.3.1 SS (Semantic Similarity)

SS uses the pre-trained model Japanese BERT to vectorize the text, and compares the text before and after augmentation by cosine similarity. The first token output from the model, CLS, is used to vectorize the text. The equation1 shows the cosine similarity used in this experiment.

$$Cos(V_o, V_a) = \frac{V_o \cdot V_a}{||V_o||\,||V_a||} \tag{1}$$

V_o : Original text vector

V_a : Augmented text vector

3.3.2 DTC (Degree of Text Change)

DTC uses BLEU(Papineni et al., 2002), a method of machine translation that evaluates translation results by comparing the translated text with the correct text using word N-grams. BLEU is characterized by the fact that the closer the translated text and the correct text are, the higher the score. In this experiment, we use BLEU provided in the NLTK(NLTK) library by default.

3.3.3 WA (Word Attention)

Word attention uses Transformer's self-attention mechanism to automatically evaluate the relationships between input data and dynamically represent the words of interest in the text. The Figure3 shows an example of the degree of attention to a text when English text is used as input data, represented by the intensity of the color. Comparing each sentence, the attention of the text along the basic syntax is the same, but the addition or replacement of a word causes a change in attention. In this research, we use a pre-trained model of Japanese BERT to extract the attention of each word from the text and compare the text before and after the augmentation. MeCab was used for data augmentation, but WA used tokenizer for segmentation.

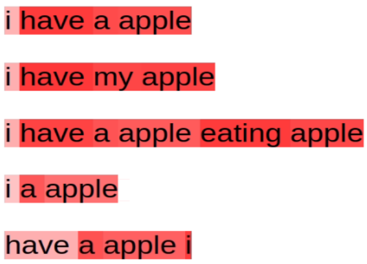

Figure 3: Example of Attention. The topmost text is the text before augmentation, and the following text is the augmented text. Colored markers indicate the degree of attention to a word.

3.4 Filtering by Evaluation Indicators

The augmented text is evaluated based on SS, DTC, and WA, and filtered by determining the respective threshold values to create the best set of augmented data for the training data. The threshold for creating optimal training data is determined by the learning accuracy obtained in training based on training data created using various combinations of threshold values for each evaluation indicator. Training accuracy refers to the percentage of correct responses when emotional polarity label classification is performed on test data. The reason for this is that we believe that the quality of the augmented data itself should be evaluated based on the learning accuracy. However, if the threshold value is set so that only those with high scores are retained in order to improve learning accuracy, a significant increase in the number of augmented data cannot be expected. Since a certain amount of data increase is necessary to improve learning accuracy, the number of training data after augmentation should also be an evaluation criterion for data augmentation. Therefore, in this research, the number of text data after augmentation is also used as a measure of data augmentation optimization, considering the balance between learning accuracy and the number of data.

3.5 Learning Model

The emotion classification model in this experiment is trained by fine-tuning BERT (Bidirectional Encoder Representations from Transformers)(Devlin et al., 2018) label classification. Fine tuning involves inputting text into the model and adjusting parameters to minimize loss between output and labels. In this experiment, we use a pre-trained Japanese language model from Tohoku University(Tohoku University) as the tokenizer and model, and evaluate the performance of emotion label classification under the conditions in the Table1. Early-Stopping means that learning is terminated if the loss in three consecutive epochs is not improved.

Model	Tohoku Uni. BERT
Tokenizer	Tohoku Uni. BERT
Learning rate	1e-5
Epoch	10
Early-Stopping	3

Table 1: Learning Environment. The learning model and parameters are shown.

4 Experiment

In this section, we present the experiment, results, discussion, and issues.

4.1 Data Augmentation

Data augmentation is performed on the training data of the WRIME corpus, and the validation data and test data are used for training without modification. For a single text, EDA generates two texts from each of the four types of text manipulation. Then, the augmented text set is the set of eight augmented texts plus the original text, from which the text identical to the original text is deleted. The augmented text is given the same label as the source text. This data augmentation process was performed on the training data. The Table2 shows examples of data augmentation with Japanese displayed in romaji.

4.2 Evaluation

SS, DTC, and WA are used to compare the text before and after data augmentation. The Table2 shows an example of the comparison: “None” in Operation indicates the text before augmentation, and “Other” indicates the text after augmentation.

4.2.1 Calculation of Word Attention Change

Table3 shows an example of how each word’s attention is noted when determining the WA. From top to bottom, it shows the original text, SR, SI, WS, and WD. Words in the original and augmented text are assigned corresponding numbers. The attention of each word represents the degree to which the model pays attention to the word, ranging from 0 to 1. Therefore, by comparing the words before and after the augmentation, the WA that the text possesses is obtained. In this experiment, the AttentionalChangeScore (ACS) is used to obtain the evaluation value by WA. However, some augmentation methods do not correspond to certain words, resulting in differences in the way WA is obtained for each augmentation method.

$$ACS = \sum_{n}(Attn_{o,n} - Attn_{a,n}) \tag{2}$$

$Attn_{o,n}$: Attention value of word$_n$ in the original text

$Attn_{a,n}$: Attention value of word$_n$ in the augmented text

The following sections describe how to obtain WA for each augmentation method.

- Synonym Replacement (SR) : In obtaining the evaluation value, the synonym-substituted word is compared with the original word.
- Synonym Insertion (SI) : To see the impact of the inserted words, the evaluation values are obtained without using the inserted words.
- Word Swap (WS) : The evaluation value is calculated from the corresponding words before and after the augmentation.

Text	Operation	SS	DTC	WA
yana kisetsu ga ki ta na xa ⋯ (The bad season is here⋯)	None	-	-	-
yana season ga ki ta na xa ⋯	SR	0.9743	0.5946	0.2544
yana kisetsu season ga ki ta na xa ⋯	SI	0.9857	0.6102	0.0759
kisetsu ga ki ta na xa ⋯ *yana*	WS	0.9858	0.7652	0.2041
kisetsu ga ki ta na xa ⋯	WD	0.9712	0.6803	0.0899

Table 2: Example of data augmentation. From left to right, the Japanese text in romaji, the operation, and the evaluation value by each evaluation indicators are shown.

Original Text

ID	1	2	3	4	5	6	7	8	9
Word	*ya*	*na*	*kisetsu*	*ga*	*ki*	*ta*	*na*	*xa*	⋯
Attention	0.709	0.517	0.482	0.667	0.732	0.558	0.708	0.437	0.687

Synonym Replacement

ID	1	2	3	4	5	6	7	8	9
Word	*ya*	*na*	*season*	*ga*	*ki*	*ta*	*na*	*xa*	⋯
Attention	0.741	0.485	0.506	0.713	0.768	0.624	0.722	0.539	0.700

Synonym Insertion

ID	1	2	3	10	4	5	6	7	8	9
Word	*ya*	*na*	*kisetsu*	*season*	*ga*	*ki*	*ta*	*na*	*xa*	⋯
Attention	0.746	0.508	0.306	0.490	0.687	0.749	0.609	0.689	0.494	0.682

Word Swap

ID	3	4	5	6	7	8	9	1	2
Word	*kisetsu*	*ga*	*ki*	*ta*	*na*	*xa*	⋯	*ya*	*na*
Attention	0.383	0.581	0.633	0.511	0.676	0.413	0.618	0.803	0.657

Word Deletion

ID	3	4	5	6	7	8	9
Word	*kisetsu*	*ga*	*ki*	*ta*	*na*	*xa*	⋯
Attention	0.428	0.622	0.645	0.559	0.713	0.455	0.714

Table 3: Examples of WA. The ID of each word, the word in the text, and the word's attention.

- Word Deletion (WD) : To see the impact of the deleted words, the evaluation values are obtained without using the deleted words.

4.2.2 Thresholds for Evaluation Indexes

The threshold values were determined based on the distribution of evaluation values for the augmented data in the Figure4, and the Table7 shows the threshold values used. From the Figure4, SS indicates that the higher the evaluation value, the more the compared texts have the same meaning. Therefore, the threshold was set within this range to ensure that the meaning does not change significantly from the original text, and because the augmented data is biased in the range of 0.9 to 1.0. In DTC, a higher evaluation value indicates that the compared texts have identical words and word sequences. Therefore, we excluded from the augmented data texts that are identical to the original texts, and since the augmented data is biased in the range of 0.5 to 0.9, we set the threshold value within this range, taking into account the number of augmented data. In WA, the closer the evaluation value is to 0, the more it indicates that the compared texts are the same in the noted parts. Therefore, the threshold was set within this range, taking into account the number of augmented data, since the text being compared was different from the original text and the augmented data was biased in the range from -0.5 to +0.5. The Table4 shows the filtering conditions by threshold value.

4.3 Filetering Example

The augmented text to be filtered using each evaluation value is shown in the following Table5. The first text from the top is the text that we want to

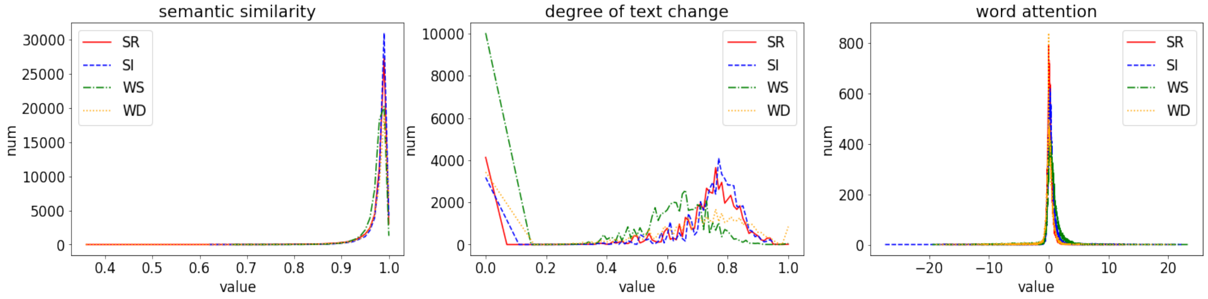

Figure 4: Distribution of evaluation indices for the augmented training data.The x-axis indicates the threshold of the evaluation value, and the y-axis indicates the amount of data.

SS	x ≥ TH
DTC	x ≤ TH
WA	x ≤ TH_{-} or TH_{+} ≤ x

Table 4: Threshold setting condition. Let x denote each evaluation value and TH denote the threshold value.

Augmented Text	SS	DTC	WA
yana season ga ki ta na xa ···	T	T	T
yana kisetsu shun ga ki ta na xa ···	T	T	F
zisetsu yana kisetsu ga ki ta na xa ···	T	F	T
yana kisetsu ta ki ga na xa ···	F	T	F

Table 5: Example of filtering.T means true to be retained, F means false to be removed.

keep the most because we believe that its meaning is similar to the text before the augmentation, the text has changed, and the model treats it as a different text. The second text has a similar meaning and textual changes. However, it is possible that the model treats it as the same text as the pre-augmentation text, so it is removed by filtering. In this way, filtering is performed when SS, DTC, or WA is False, and the evaluation index is used on the extended text to select the extended text suitable for training the model.

4.4 Learning

We perform experiments using the training data from the WRIME corpus, the validation data, and the test data. The Table6 shows the breakdown of the number of data in each category. We also check the change in learning accuracy by expanding data only on the training data, and using the same validation and test data for all training. The training accuracy is the percentage of correct answers to the model trained on the training data using the test data.

Type	Size
Train	30,000
Valid	2,500
Test	2,500

Table 6: WRIME Corpus. Training data, validation data, and test data composition.

4.4.1 Learning Results without Filtering

The WRIME corpus is used as the original data, and data augmentation to the training data is used as the augmented training data. Using these data, we compare the learning accuracy of emotional polarity label classification. The Table7 shows the data size and learning accuracy of the training data before and after augmentation.

4.5 Learning Results with Filtering

We evaluated the augmented training data using SS, DTC, and WA, and created new augmented training data using the threshold values. Then, we compare the learning accuracy of emotional polarity label classification. The Table7 shows the data size and learning accuracies for the augmented training data, applying the evaluation indicators to the augmented training data.

Next, we compared the learning accuracy of emotional polarity label classification when combining SS, DTC, and WA thresholds. The Table7 shows the data size and learning accuracies for the optimal thresholds, taking into account the respective evaluation values and the number of augmented data.

5 Discussion and Issues

In this section, we compare the learning of classification of emotional polarity labels using the training data after data augmentation based on the experimental results in the previous section with

Evaluation Indicators			Size	5-Labels		3-Labels	
SS	DTC	WA		Subj.	Obj.	Subj.	Obj.
without Filtering							
Original Training Data			30,000	0.391	0.566	0.616	0.697
Augmented Training Data			243,788	0.388	0.560	0.576	**0.704**
with Filtering							
0.9	-	-	240,001	0.392	0.563	0.575	**0.708**
0.95	-	-	226,638	0.374	0.554	0.556	0.695
0.98	-	-	172,791	0.379	0.564	0.585	**0.704**
-	0.7	-	124,452	**0.410**	**0.570**	**0.631**	0.695
-	0.6	-	84,219	**0.413**	**0.572**	**0.641**	0.686
-	0.5	-	63,047	**0.414**	**0.574**	**0.626**	0.696
-	-	-0	87,466	0.392	**0.576**	0.618	**0.708**
-	-	+0	185,751	0.373	0.562	0.586	**0.703**
-	-	0.5	97,736	**0.404**	0.554	**0.633**	0.699
Filtering by multiplying two Evaluation Indicators							
0.99	0.6	-	61,823	**0.411**	0.569	**0.650**	**0.705**
0.95	-	+0	174,342	**0.411**	0.561	0.611	**0.701**
-	0.6	0.5	45,930	**0.424**	**0.573**	**0.653**	**0.703**
Filtering by multiplying three Evaluation Indicators							
0.95	0.7	0.5	60,096	0.394	**0.572**	**0.625**	**0.705**

Table 7: Experimental Results. From left to right, thresholds for each evaluation indicators, data size, and learning accuracies for the five and three sentiment polarity labels. From top to bottom, training on the original data, training on the augmented data, and training with thresholds applied to the augmented data.

the learning by filtering process using the evaluation indicators, and discuss the issues involved.

5.1 Discussion of Unfiltered Learning Results

The results of the comparison of the learning accuracy of the emotional polarity label classification by BERT showed that the learning accuracy of the five subjective emotional polarity labels was 0.391 using the original training data and 0.388 using the augmented training data, and no improvement in accuracy could be confirmed. Similarly, no clear improvement in accuracy was observed for the three subjective emotion polarity labels. However, for the three objective emotion polarity labels, the training accuracy using the augmented training data was 0.704, while the accuracy using the original training data was 0.697, showing a slight improvement in accuracy. The reason for the lack of improvement in learning accuracy is that subjective labels are more affected by differences in the tendency of each writer to assign labels than are objective labels, which use the average of labels assigned by multiple readers. Therefore, subjective labels are more susceptible to the influence of slight changes in meaning due to data augmentation. As a result, it is thought that data that reduces the learning accuracy may have been mixed in. For example, since different writers have different labeling tendencies, we believe that increasing the amount of training data created by multiple writers will make it easier to misclassify data created by writers with different tendencies.

5.2 Discussion of Filtered Learning Results

The Table7 shows that the learning accuracy was slightly improved in label classification of emotional polarity for the augmented training data filtered by each evaluation indicators, compared to that using the original training data. However, there were cases in which the learning accuracy did not improve, such as when the threshold value increased the percentage of correct subjective labels and decreased the percentage of correct objective labels. In terms of each evaluation indicator, DTC showed a clear improvement in learning accuracy, but filtering by SS showed no improvement in learning accuracy. The filtering by WA also showed no clear improvement in learning accuracy.We believe

Augmented Text	SS	DTC	WA
yana season ga ki ta na xa ⋯	0.97	0.59	-0.25
yana yoki ga ki ta na xa ⋯	0.96	0.59	-0.01
yana kisetsu ga kita ⋯	0.93	0.70	0.29

Table 8: Example of text you do not want to filter.

that the reason for the lack of improvement in learning accuracy is that SS and WA are dependent on model performance, and that the evaluation values may not be output correctly.

As an example, the first and second augmented texts from the Table8 are texts generated by synonym replacement.The low SS is due to the model not learning enough of the replacement words, which we consider to be the reason for the low similarity. In addition, the WA may be lower depending on the location of the replacement. The third text is the text generated by word deletion. It is considered to have the same meaning to people as before the augmentation, but the evaluation of the model shows a low SS and is not considered similar to before the augmentation.

In order to perform the evaluation correctly, we believe it is necessary to construct a model-independent method and a model suited to the data set.

The table shows that the accuracy in label classification of emotional polarity of the augmented training data filtered by a combination of each evaluation indicator was better than that of the training data filtered by each evaluation indicator. In particular, the training data filtered using DTC and WA shows the most improvement in label classification. We believe that this means that only high quality data can be used for training data by filtering. However, the training data size is greatly reduced when filtering for the combined evaluation indicators, and we believe that the optimal augmentation method is not being used to generate the data. Therefore, it is necessary to investigate a suitable data augmentation method for Japanese texts.

6 Conclusion

In this research, we aimed to improve the learning accuracy of label classification of sentiment polarity by augmenting Japanese text data with data augmentation, evaluating the augmented text, and filtering the post-extension training data. In particular, by using SS, DTC, and WA for the evaluation of augmented text, we were able to generate data suitable for learning. As a result, learning accuracy of label classification was improved by using augmented training data filtered by a combination of SS, DTC, and WA. We found that there are issues such as Japanese text augmentation methods and model dependence between SS and WA. In order to solve these issues, we would like to investigate augmentation methods and evaluation methods, and examine whether appropriate data is generated by data augmentation.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Numbers JP20K12027, JKA and its promotion funds from KEIRIN RACE.

References

George A. and Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, page 38(11):39–41.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *Computation and Language*, arXiv:1810.04805. Version 2.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. *Proceedings of the Third International Workshop on Paraphrasing (IWP2005).*

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. *In Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 168–177.

Tomoyuki Kajiwara, Chenhui Chu, Noriko Takemura, Yuta Nakashima, and Hajime Nagahara. 2021. Wrime: A new dataset for emotional intensity estimation with subjective and objective annotations. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (In Japanese).*

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to japanese morphological analysis. *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (In Japanese)*, pages 230–237.

Xin Li and Dan Roth. 2002. Learning question classifiers. *In Proceedings of the 19th International Conference on Computational Linguistics*, 1:1–7.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A sick cure for the evaluation of compositional distributional semantic models. *Proceedings* *ofthe Ninth International Conference on Language Resources and Evaluation (LREC' 14).*

NLTK. Bleu of nltk. *https://github.com/nltk/nltk.*

Itsuki Okimura, Makoto Kawano, Machel Reid, and Yutaka Matsuo. 2022. Analyzing the impact of data augmentation on performance improvement in natural language. *The 36th Annual Conference of the japanese Society for Artificial Inteligence (In Japanese).*

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. *In Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Richard Socher, Alex Perelygin, Jean Wu, Jason, Chuang, Christopher Manning, Andrew Ng, and Christopher Potts. 2013. Parsing with compositional vector grammars. *In EMNLP*.

Tohoku University. Pretrained japanese bert models. *https://github.com/cl-tohoku/bert-japanese*.

Haruto Uda, Kazuyuki Matsumoto, Minoru Yoshida, and Kenji Kita. 2023. Investigation on accuracy improvement of emotion classificaton based on text data. *The 37th Annual Conference of the japanese Society for Artificial Inteligence (In Japanese).*

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30:6000–6010.

Jason Wei and Kai Zou. 2019. Eda:easy data augmentation techniques for boosting performance on text classification tasks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388.

Ichiro Yamada, Jong hoon Oh, Kentaro Torisawa, Wataru Kuroda, Jun ichi Kazama, and Maki Murata. 2010. study of term addition to japanese wordnet using wikipedia. *The 16th Annual Meeting of the Association for Natural Language Processing (In Japanese).*

Toshiteru Yamada, Syota Harada, and Seiichi Uchida. 2022. Adaptive selection of data augmentation with transformer. *Proceedings of the 2022 Kyushu Section Joint Convention of the Institutes of Electrical and Information Engineers (The 75th Joint Convention) (In Japanese).*

A Example Appendix

The text of Table2, Table3, Table5 and Table 8 are shown in Japanese.

text	Operation	semantic similarity	degree of text change	word attention
やな季節が来たな⋯ (The bad season is here⋯)	None	-	-	-
やなシーズンが来たなぁ⋯	SR	0.9743	0.5946	0.2544
やな季節シーズンが来たなぁ⋯	SI	0.9857	0.6102	0.0759
季節が来たなぁ⋯やな	WS	0.9858	0.7652	0.2041
季節が来たなぁ⋯	WD	0,9712	0.6803	0.0899

Table 9: Example of data augmentation in Japanese. From left to right, the Japanese text in romaji, the operation, and the evaluation value by each evaluation indicators are shown.

Original Text

ID	1	2	3	4	5	6	7	8	9
word	や	な	季節	が	来	た	な	ぁ	⋯
attention	0.709	0.517	0.482	0.667	0.732	0.558	0.708	0.437	0.687

Synonym Replacement

ID	1	2	3	4	5	6	7	8	9
word	や	な	シーズン	が	来	た	な	ぁ	⋯
attention	0.741	0.485	0.506	0.713	0.768	0.624	0.722	0.539	0.700

Synonym Insertion

ID	1	2	3	10	4	5	6	7	8	9
word	や	な	季節	シーズン	が	来	た	な	ぁ	⋯
attention	0.746	0.508	0.306	0.490	0.687	0.749	0.609	0.689	0.494	0.682

Word Swap

ID	3	4	5	6	7	8	9	1	2
word	季節	が	来	た	な	ぁ	⋯	や	な
attention	0.383	0.581	0.633	0.511	0.676	0.413	0.618	0.803	0.657

Word Deletion

ID	3	4	5	6	7	8	9
word	季節	が	来	た	な	ぁ	⋯
attention	0.428	0.622	0.645	0.559	0.713	0.455	0.714

Table 10: Examples of WA in Japanese. The ID of each word, the word in the text, and the word's attention.

Augmented Text	SS	DTC	WA
やなシーズンが来たなぁ⋯	T	T	T
やな季節旬が来たなぁ⋯	T	T	F
時節やな季節が来たなぁ⋯	T	F	T
やな季節た来がなぁ⋯	F	T	F

Table 11: Example of filtering in Japanese. T means true to be retained, F means false to be removed.

Augmeted Text	SS	DTC	WA
やなシーズンが来たなぁ⋯	0.97	0.59	-0.25
やな陽気が来たなぁ⋯	0.96	0.59	-0.01
やな季節が来た⋯	0.93	0.70	0.29

Table 12: Example of text you do not want to filter in Japanese.