algorithms

MDPI

Article

Data Augmentation Methods for Enhancing Robustness in Text Classification Tasks

Huidong Tang *, Sayaka Kamei and Yasuhiko Morimoto *

Graduate School of Advanced Science and Engineering, Hiroshima University, Kagamiyama 1-7-1, Higashi-Hiroshima 739-8521, Japan

* Correspondence: d216083@hiroshima-u.ac.jp (H.T.); morimo@hiroshima-u.ac.jp (Y.M.); Tel.: +81-82-424-5579 (Y.M.)

Abstract: Text classification is widely studied in natural language processing (NLP). Deep learning models, including large pre-trained models like BERT and DistilBERT, have achieved impressive results in text classification tasks. However, these models' robustness against adversarial attacks remains an area of concern. To address this concern, we propose three data augmentation methods to improve the robustness of such pre-trained models. We evaluated our methods on four text classification datasets by fine-tuning DistilBERT on the augmented datasets and exposing the resulting models to adversarial attacks to evaluate their robustness. In addition to enhancing the robustness, our proposed methods can improve the accuracy and F1-score on three datasets. We also conducted comparison experiments with two existing data augmentation methods. We found that one of our proposed methods demonstrates a similar improvement in terms of performance, but all demonstrate a superior robustness improvement.

Keywords: artificial intelligence; natural language processing; text classification; data augmentation; robustness improvement

check for updates

Citation: Tang, H.; Kamei, S.; Morimoto, Y. Data Augmentation Methods for Enhancing Robustness in Text Classification Tasks. *Algorithms* **2023**, *16*, 59. https://doi.org/10.3390/a16010059

Academic Editors: Ole-Christoffer Granmo, Lei Jiao, Vladimir Zadorozhny and Morten Goodwin

Received: 14 December 2022
Revised: 9 January 2023
Accepted: 12 January 2023
Published: 16 January 2023

1. Introduction

Text classification is widely studied in the field of natural language processing (NLP), encompassing tasks such as spam detection [1], topic classification [2], and sentiment classification [3]. Deep learning models, including large pre-trained models like BERT [4], have achieved impressive results in text classification tasks. Furthermore, distilled pre-trained models like DistilBERT [5] allow for deploying these large models on edge devices. While these smaller models maintain a strong performance, their robustness against adversarial attacks remains an area of concern [6–8] as such attacks have been shown to deceive these pre-trained models potentially. In this work, we define robustness as the ability of a model to resist adversarial attacks.

Previous studies [6,8,9] have primarily focused on using adversarial examples to enhance the robustness of deep learning models. While these methods can be effective, they require the presence of a victim model to generate adversarial examples. In contrast, data augmentation techniques for text classification, which do not require a victim model, have been proposed in several studies [10–17]. These methods are generally designed to improve the model's performance rather than robustness and often involve replacing words with similar words. While these methods are effective in improving performance, they need to be improved in their ability to generate text with diverse semantics and lexical variety, which can result in a lack of robustness.

Motivated by the potential for data augmentation to improve the robustness, we propose three methods going beyond replacing words with similar words. Our Cognate-based, Antonym-based, and Antipode-based methods are designed to enhance the robustness of pre-trained models. Cognate-based methods utilize synonyms and words with similar sentiments to increase the lexical diversity, while Antonym-based methods incorporate antonyms to provide a semantic variation. Antipode-based methods, on the other hand, combine antonyms and words with opposite sentiments to create both semantic and lexical diversity, leading to an increased robustness. Overall, these approaches enhance the robustness of the generated text by incorporating a greater range of semantic and lexical variation.

Pure Antonym-based data augmentation methods have been previously used for sentiment analysis, but the two other data augmentation methods we propose are entirely novel. These methods combine synonyms and similar sentiment words (Cognate-based) or antonyms and sentiment words (Antipode-based).

This study defines robustness as accuracy under attack and the attack success rate. Accuracy under attack is the accuracy of a target model under an adversarial attack, measured as the number of samples that resist the attack out of the total number of samples. The attack success rate is the proportion of adversarial attacks that successfully change the prediction of a target model from correct to incorrect. These data augmentation processes and victim models are independent.

Our proposed methods can potentially be utilized not only for the mentioned text classification tasks but also for tasks such as topic segmentation [18] and authorship attribution [19] that place a high value on robustness due to their semantic lexical diversity and potential for counterfeiting.

We evaluated our proposed data augmentation methods on the AG-News [20], TREC [21,22], SUBJ [23], and SMS-Spam [24] datasets. We fine-tuned DistilBERT [5] on the augmented datasets and then exposed the resulting models to adversarial attacks to evaluate their robustness. In addition to enhancing the robustness, our proposed methods improved the accuracy and F1-score on three datasets. We also conducted comparison experiments with two existing data augmentation methods: EDA [10] and CheckList [7]. We found that one of our proposed methods performs a similar improvement in terms of the performance, but all demonstrate a superior robustness improvement. Our contributions can be summarized as follows:

1. To the best of our knowledge, we are the first to apply data augmentation techniques to improve the robustness.
2. We propose two novel data augmentation methods for text classification: Cognate-based methods, which combine synonyms and similar sentiment words, and Antipode-based methods, which combine antonyms and sentiment words.
3. We conducted experiments on the AG-News [20], TREC [21,22], SUBJ [23], and SMS-Spam [24] datasets and demonstrated the effectiveness of our methods.
4. We also conducted comparison experiments with two existing data augmentation methods and found that one of our proposed methods performs similarly but demonstrates a superior robustness.

2. Related Work

This section reviews the current research on text adversarial example generation and text data augmentation for natural language processing (NLP) models. The focus of the investigation on adversarial examples is the enhancement of the model's robustness, while the use of data augmentation is geared towards improving the accuracy.

2.1. Text Adversarial Examples Generation

Various research has been devoted to generating adversarial text examples in recent years due to the continued vulnerability of deep learning models to adversarial attacks. Some of the notable approaches include TextFooler [6], which employs a combination of word importance ranking and synonym-based word transformation, and CLARE [8], which uses a pre-trained masked language model and executes three perturbation actions (replace, insert, and merge) to generate contextualized adversarial examples. Another approach, PWWS [9], utilizes synonym substitution to generate adversarial examples, automatically determining the most effective replacement.

While these approaches have demonstrated success in improving the robustness of victim models, it is essential to note that the adversarial example generation process and victim model are not independent.

2.2. Text Data Augmentation

There has been a proliferation of research on data augmentation methods for deep learning models in recent years, intending to improve performance metrics such as the accuracy and F1-score. Examples of such methods include EDA [10], which employs a combination of random synonym replacement, random word deletion, random word position swap, and random synonym insertion, and AEDA [11], which randomly inserts punctuation marks into the original text. Other approaches include Data Boost [12], which utilizes a large-scale pre-trained language model with reinforcement learning optimization, and LAMBDA [13], which also uses a large-scale pre-trained model as a generator, fine-tuning it and training an additional classifier using identical datasets. The augmented data are then filtered using the classifier to obtain a high-quality dataset. UDA [14] applies data augmentation methods from supervised learning to semi-supervised learning, and Contextual Augmentation [15] employs a label-conditional language model to predict paradigmatic relation-based word replacement. In [16], dependency trees break sentences into smaller pieces and change the order of words for data augmentation in low-resource languages. [17] develops methods to find suitable text modification operations for data augmentation automatically.

These methods are generally designed to improve the model's performance rather than robustness and often involve replacing words with similar words. While these methods effectively improve the performance, they are limited in their ability to generate text with diverse semantics and lexical variety, which can result in a lack of robustness.

3. Materials and Methods

This section provides a detailed description of the datasets utilized in our study and describes our proposed methods.

3.1. Datasets

In order to evaluate the effectiveness of our proposed data augmentation methods, we conducted experiments on a variety of datasets, including AG-News [20], TREC [21,22], SUBJ [23], and SMS-Spam [24]. To ensure the robustness of our results, we randomly sampled and shuffled the data from these datasets, as described in Table 1, which lists the tasks and sizes of each dataset.

Table 1. Datasets Description.

	Tasks	# of Classes	Training Size	Evaluation Size	Test Size
AG-News	Topic Classification	Four	5000	500	500
TREC	Question Classification	Six	4500	500	500
SUBJ	Movie Description Classification	Two	5000	500	500
SMS-Spam	Spam Detection	Two	4500	500	500

3.2. Methods

Our proposed data augmentation methods aim to replace a randomly selected subset of words in a given text with substitutes from predetermined lists of words. We follow two constraints while selecting words in the given text for word replacement; we do not select words that have already been replaced or stop words in the text.

Our Cognate-based, Antonym-based, and Antipode-based approaches aim to generate these lists. Specifically, we randomly select a percentage (n%) of words from the text and replace them with words from the generated lists while adhering to certain constraints, particularly regarding part-of-speech (POS). We describe these three proposed data augmentation methods in further detail below:

- Cognate-based methods generate a list of substitutes by compiling synonyms and words with the similar sentiment using WordNet [25] and SenticNet [26]. In SenticNet [26], sentiment-related words are rated based on sensitivity, attitude, temper, and introspection scores, representing twenty-four basic emotions (these twenty-four basic emotions are ecstasy, joy, contentment, melancholy, sadness, grief, bliss, calmness, serenity, annoyance, anger, rage, delight, pleasantness, acceptance, dislike, disgust, loathing, enthusiasm, eagerness, responsiveness, anxiety, fear, and terror). For instance, a high introspection score indicates intense ecstasy, while a low introspection score indicates intense grief. Similar sentimental words must have the identical four scores as the target word to be included in the substitute list.
- Antonym-based methods generate a list of substitutes by compiling antonyms using WordNet [25].
- Antipode-based methods generate a list of substitutes by compiling antonyms and words with the opposite sentiment using WordNet [25] and SenticNet [26]. To be included in the substitute list, opposite sentimental words must have the opposite four scores of the target word.

We apply three different constraints on the substitute lists used for word replacement, resulting in three variations of our Cognate-based, Antonym-based, and Antipode-based methods, respectively:

- Cognate1 (resp. Antonym1, Antipode1): all words in the substitute list generated by the Cognate-based (resp. Antonym-based, Antipode-based) method can be used for word replacement.
- Cognate2 (resp. Antonym2, Antipode2): only words in the substitute list generated by the Cognate-based (resp. Antonym-based, Antipode-based) method with the same part-of-speech (POS) as the selected word in the given text can be used for word replacement, except that verbs can be replaced with nouns and vice versa.
- Cognate3 (resp. Antonym3, Antipode3): only words in the substitute list generated by the Cognate-based (resp. Antonym-based, Antipode-based) method with the same POS as the selected word in the given text can be used for word replacement.

The following samples from AG-News [20] are presented below, with words highlighted in bold indicating those that have been replaced in generated samples:

- Original: "Bangladesh paralysed by strikes Opposition activists have brought many towns and cities in Bangladesh to a halt, the day after 18 people died in explosions at a political rally."
- Cognate-based: "Bangladesh paralysed by strikes Opposition activists have brought many **townsfolk** and cities in Bangladesh to a halt, the day after 18 people **edema** in explosions at a political rally."
- Antonym-based: "Bangladesh paralysed by strikes Opposition activists have brought **few** towns and cities in Bangladesh to a halt, the day after 18 people died in explosions at a **nonpolitical** rally."
- Antipode-based: "Bangladesh paralysed by strikes Opposition activists have brought many towns and cities in Bangladesh to a halt, the **night** after 18 people died in explosions at a political **demobilize**."

4. Evaluation Experiments

This section outlines the experimental procedures employed for evaluation and presents the results.

4.1. Implementation

Our evaluation experiments have three main stages: data augmentation, model generation, and text attack. Specifically, we first apply data augmentation methods to the datasets three times each, generating augmented datasets. These augmented datasets are then used to fine-tune DistilBERT [5] using identical learning parameters and random seed. Next, we employ TextFooler [6] and PWWS [9] to assess the robustness of the fine-tuned models and average the results obtained from each data augmentation method. In addition, we also compared with EDA [10] and the transformation method from CheckList [7], which combines name replacement, location replacement, number alteration, and contraction/extension. These experiments are implemented using Textattack [27] and Transformers [28].

4.1.1. Data Augmentation

To augment the training dataset, we employed the data augmentation methods described in Section 3 and the EDA [10] and CheckList [7] methods for comparative purposes. All data augmentation methods were implemented according to the following protocol: we randomly select 10% of the words in the text for replacement and generate one augmented sample for each original data sample.

4.1.2. Model Generation

To generate the models, we fine-tuned DistilBERT [5] using both the original dataset and the augmented datasets produced by each data augmentation method. In order to isolate the effect of data augmentation and make the model generation process more efficient, we fixed the learning parameters and random seed for all models. We employed padding and truncation strategies to maintain a fixed input length for each dataset. The input length for AG-News [20] is fixed at 200, TREC [21,22] at 50, and both SUBJ [23] and SMS-Spam [24] at 100. We evaluated the performance of the generated models in terms of the accuracy and F1-score during the model generation process.

4.1.3. Text Attack

To assess the robustness of the generated models, we utilize TextFooler [6] and PWWS [9] (excluding the named entity adversarial swap) as attackers, which modify the test dataset input in an attempt to cause the victim model to produce incorrect predictions. To eliminate the impact of randomness on the attacks, we fix the random seed. While the input length for AG-News [20], TREC [21,22], SUBJ [23], and SMS-Spam [24] is fixed, during the model generation process, in order to maintain generalizability during actual deployment, the victim models in the text attack process use the default input length setup of Transformers [28] (i.e., no padding or truncation strategies). We evaluate the robustness in terms of the accuracy under attack and the attack success rate, which is defined as the number of samples originally predicted correctly that were forced to predict incorrectly under attack, divided by the number of samples originally predicted correctly without attack, as shown in Equation (1):

$$\mathrm{R_{AS}} = \mathrm{N_{PI}}/\mathrm{N_{OPC}}, \tag{1}$$

where $\mathrm{R_{AS}}$ is the attack success rate, $\mathrm{N_{PI}}$ is the number of samples forced to predict incorrectly under attack, and $\mathrm{N_{OPC}}$ is the number of samples originally predicted correctly without attack.

4.2. Experiment Results

The performance results, including the accuracy and F1-score, are presented in Section 4.2.1, while the robustness results, including the accuracy under attack and attack success rate, are presented in Section 4.2.2. These results are presented as the difference between the results of the data augmentation methods and the original model. Detailed performance and robustness results can be found in Tables A1–A4 in the Appendix A.

4.2.1. Performance Results

The results of the AG-News dataset [20] in Figure 1 indicate that all data augmentation methods improve the accuracy and F1-score compared to the original model. Among these methods, *Antipode3* produces the most significant improvement, with an increase of 1.87% in the accuracy and 1.91% in the F1-score.

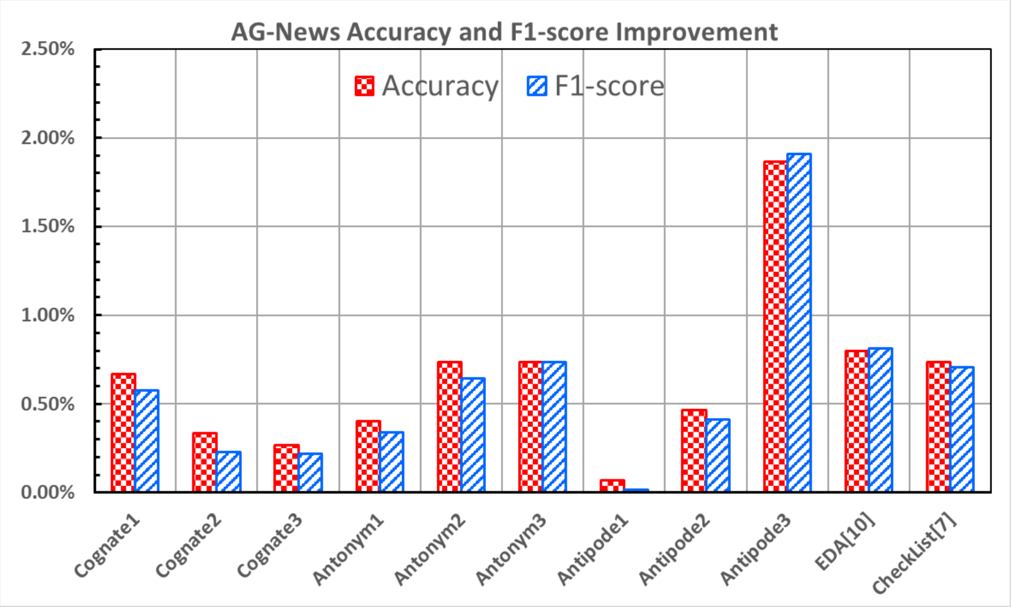

Figure 1. AG-News Performance Improvement Results.

The results of the TREC dataset [21,22] depicted in Figure 2 reveal that, compared to the original model, all data augmentation methods decrease both the accuracy and F1-score. The exception to this trend is *Antonym2*, which results in a slight increase in the accuracy (0.07%) but a decrease in the F1-score.

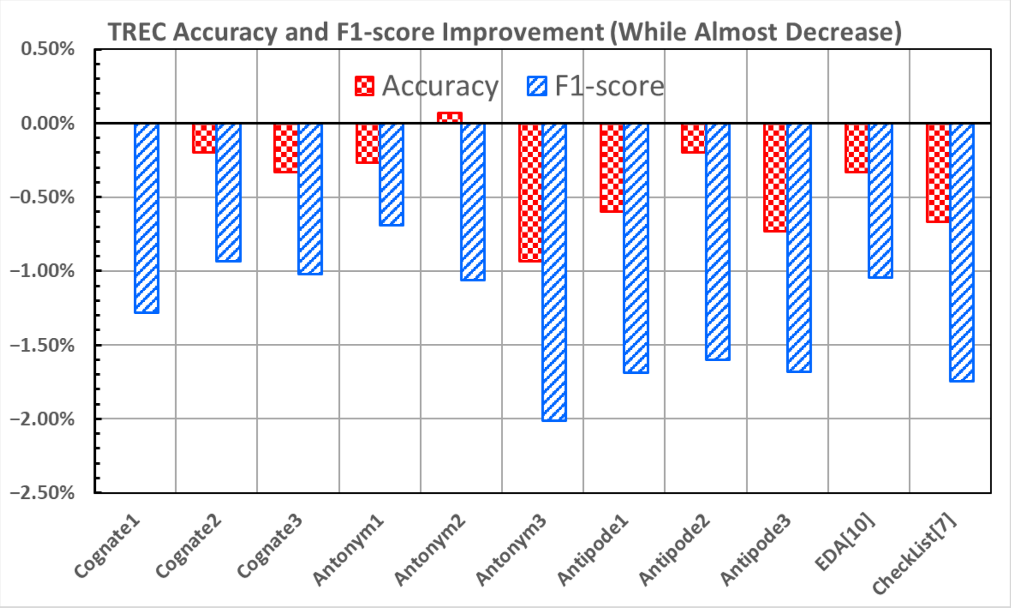

Figure 2. TREC Performance Improvement Results.

As illustrated in Figure 3, the results of the SUBJ dataset [23] demonstrate that, with some exceptions, most data augmentation methods improve the accuracy and F1-score compared to the original model. EDA [10] produces the most significant improvement of these methods, with an increase of 0.60% in both the accuracy and F1-score. *Antonym1* also shows comparable results, with an increase of 0.40% in the accuracy and 0.39% in the F1-score.

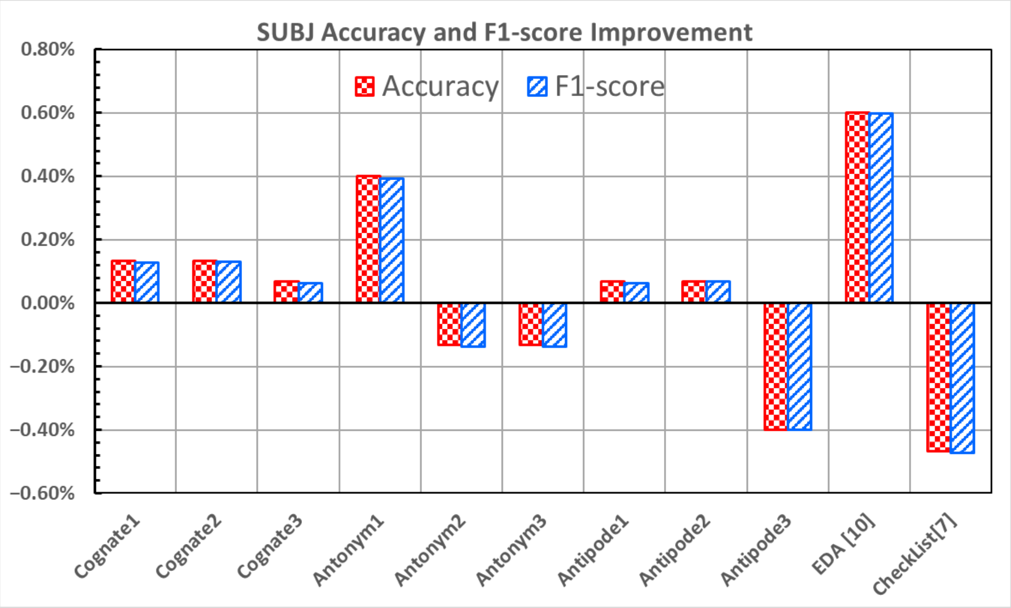

Figure 3. SUBJ Performance Improvement Results.

The results of the SMS-Spam dataset [24] depicted in Figure 4 indicate that all data augmentation methods improve the accuracy and F1-score compared to the original model. Among these methods, CheckList [7] produces the most significant improvement, with an increase of 0.53% in the accuracy and 1.22% in the F1-score. *Antipode2* also demonstrates comparable results, with an increase of 0.47% in the accuracy and 1.05% in the F1-score.

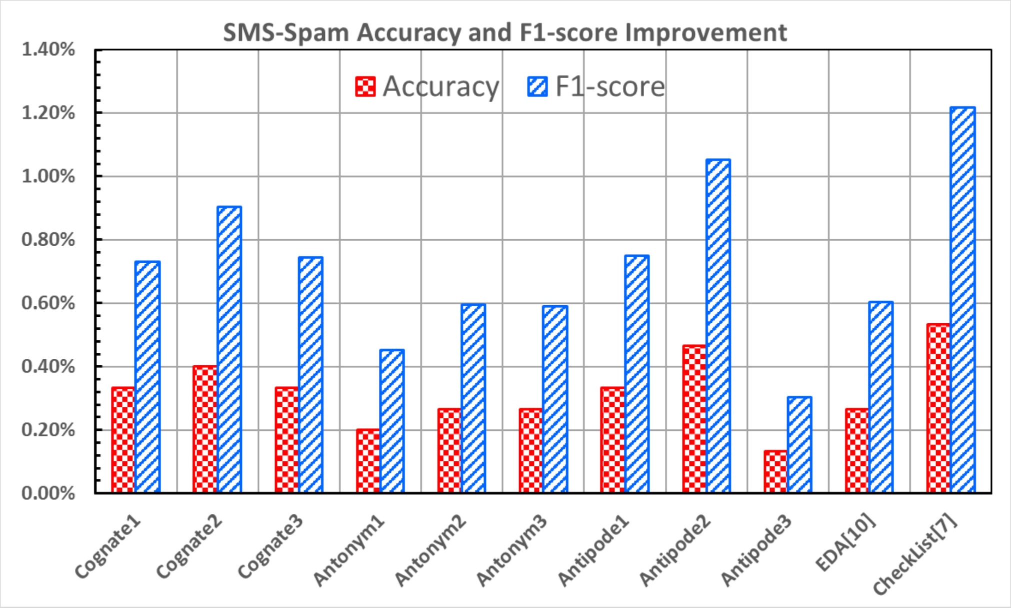

Figure 4. SMS-Spam Performance Improvement Results.

The most effective methods for improving the performance vary across datasets. While our proposed methods may only consistently outperform others on some datasets, the

Antipode-based methods demonstrate a high level of versatility and efficacy, suggesting their utility in various contexts.

4.2.2. Robustness Results

The results presented in this section demonstrate the effectiveness of our proposed data augmentation methods in improving the robustness of the models under attack by TextFooler [6] and PWWS [9]. Figures 5 and 6 show the improvement in the accuracy under attack and the decrease in the attack success rate, respectively, for the AG-News dataset [20]. *Cognate3, Antonym2,* and *Antipode2* exhibit notable increases in the accuracy under attack and decreases in the attack success rate under both attacks. In particular, *Antipode2* demonstrates a significant increase in the accuracy under attack (2.80% and 8.60% in Figure 5) and a decrease in the attack success rate (−3.00% and −9.34% in Figure 6) in both attacks.

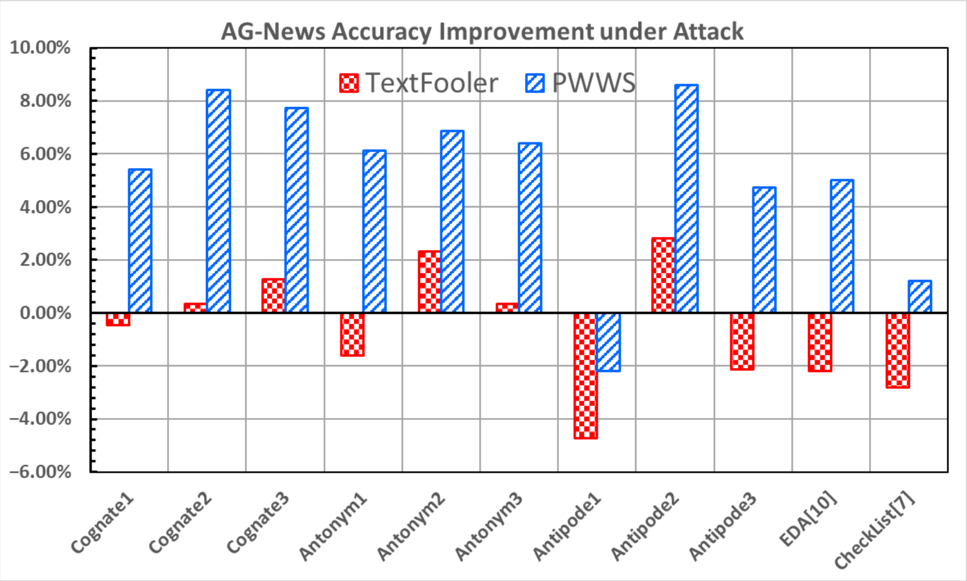

Figure 5. AG-News Accuracy Improvement Under Attack.

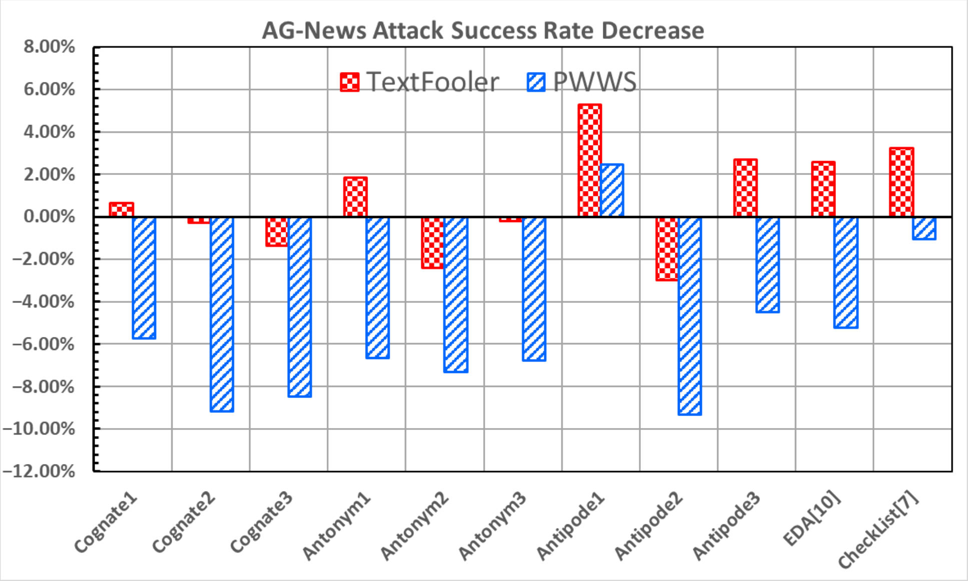

Figure 6. AG-News Attack Success Rate Decrease.

Figures 7 and 8 show the improvement in the accuracy under attack and the decrease in the attack success rate for the TREC dataset [21,22]. *Cognate2, Antonym1, Antonym2, Antipode1,* and CheckList [7] all exhibit increases in the accuracy under attack and decreases in the attack success rate under both attacks. Specifically, *Antonym1* and *Antonym2* show solid results, with increases in the accuracy under attack (1.07% and 3.80%; 1.67% and 1.27% in Figure 7) and decreases in the attack success rate (−1.19% and −4.08%; −1.70% and −1.27% in Figure 8) in both attacks.

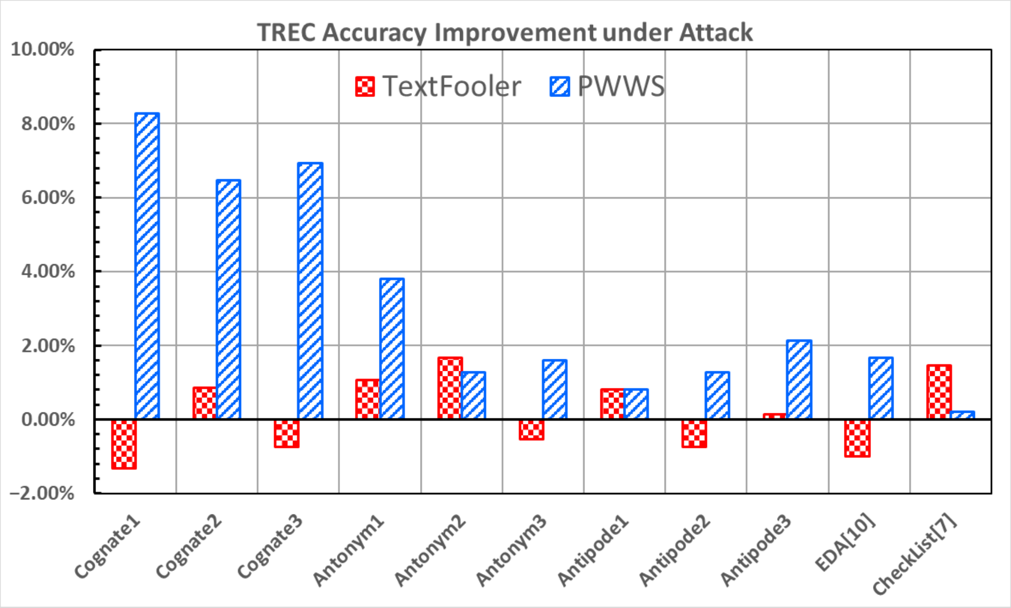

Figure 7. TREC Accuracy Improvement Under Attack.

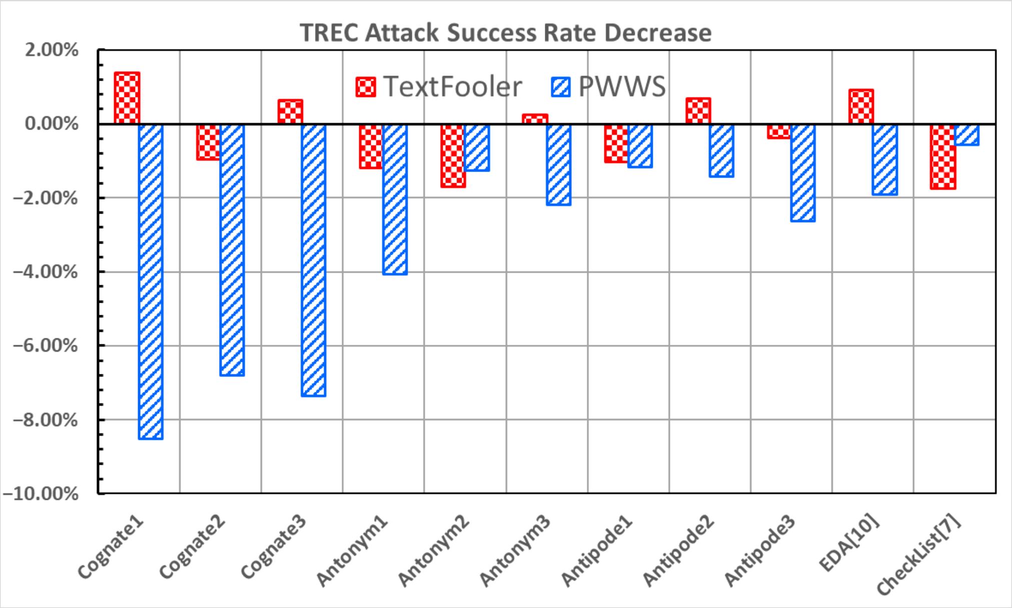

Figure 8. TREC Attack Success Rate Decrease.

Figures 9 and 10 show the improvement in the accuracy under attack and the decrease in the attack success rate, respectively, for the SUBJ dataset [23]. *Antonym2* and *Antipode1* demonstrate notable increases in the accuracy under attack (1.13% and 2.33%; 1.00% and 2.53% in Figure 9) and decreases in the attack success rate (−1.20% and −2.48%; −1.02% and −2.61% in Figure 10) under both attacks.

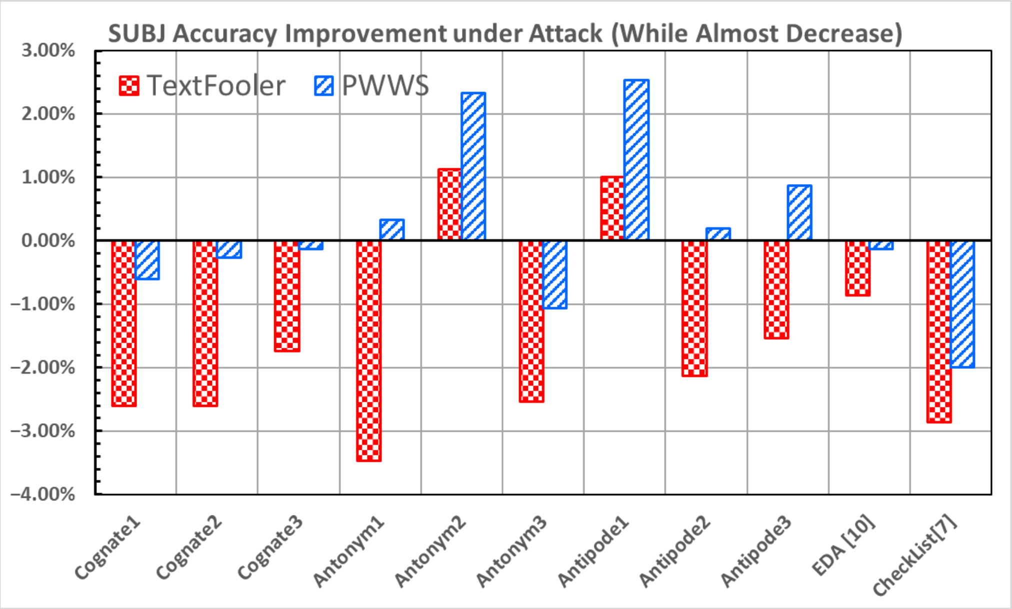

Figure 9. SUBJ Accuracy Improvement Under Attack.

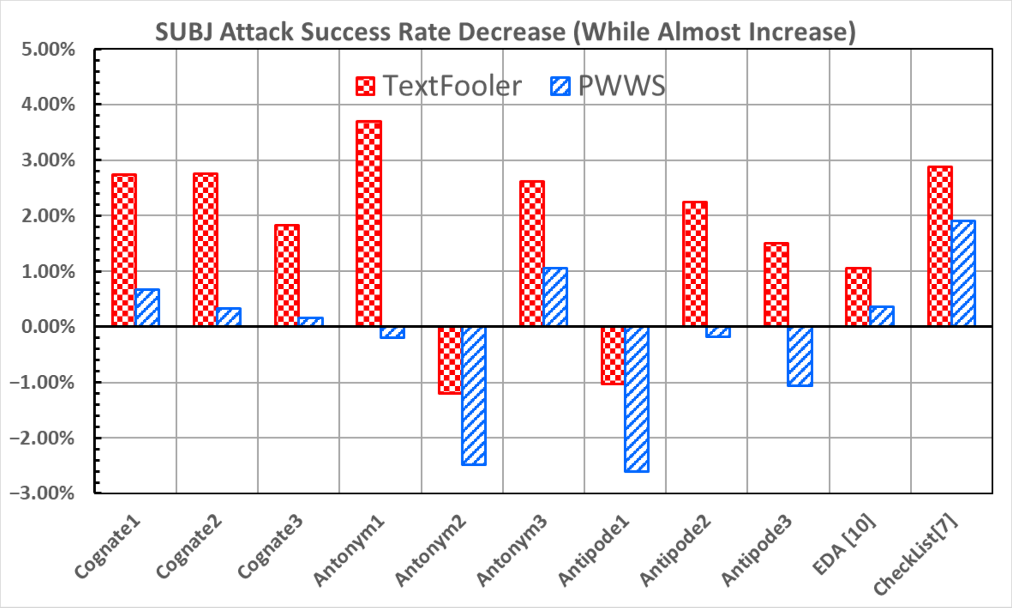

Figure 10. SUBJ Attack Success Rate Decrease.

Figures 11 and 12 show the improvement in the accuracy under attack and the decrease in the attack success rate for the SMS-Spam dataset [24]. All data augmentation methods exhibit increases in the accuracy under attack and decrease in the attack success rate under both attacks. *Cognate1* and *Antipode2* show solid results, with *Antipode2* demonstrating an impressive increase in the accuracy under attack (16.93% and 11.47% in Figure 11) and a decrease in the attack success rate (−16.68% and −11.15% in Figure 12) in both attacks. The SMS-Spam dataset's limited semantics and lexical diversity [24] may contribute to the observed results.

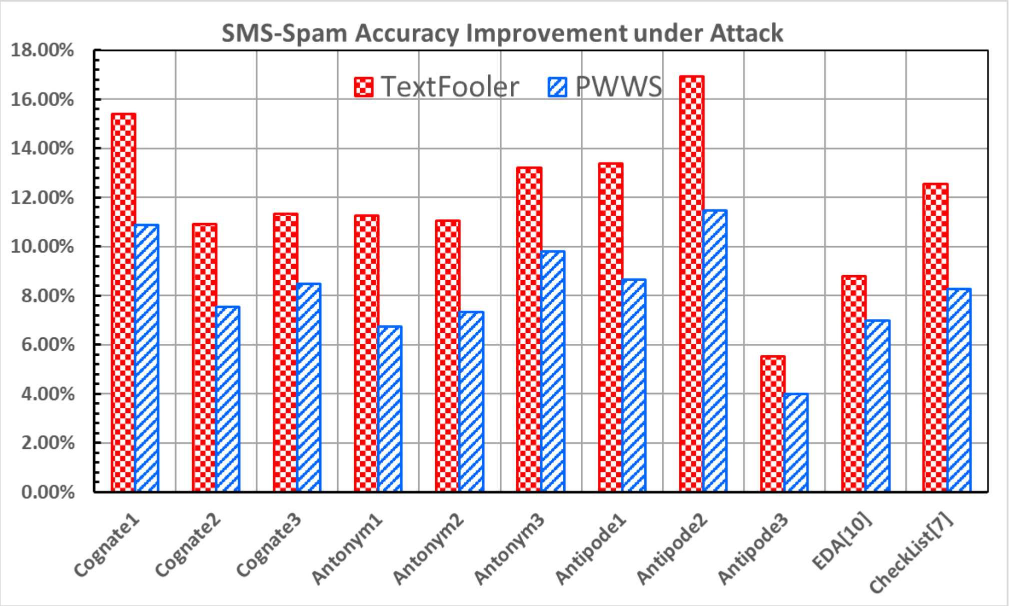

Figure 11. SMS-Spam Accuracy Improvement Under Attack.

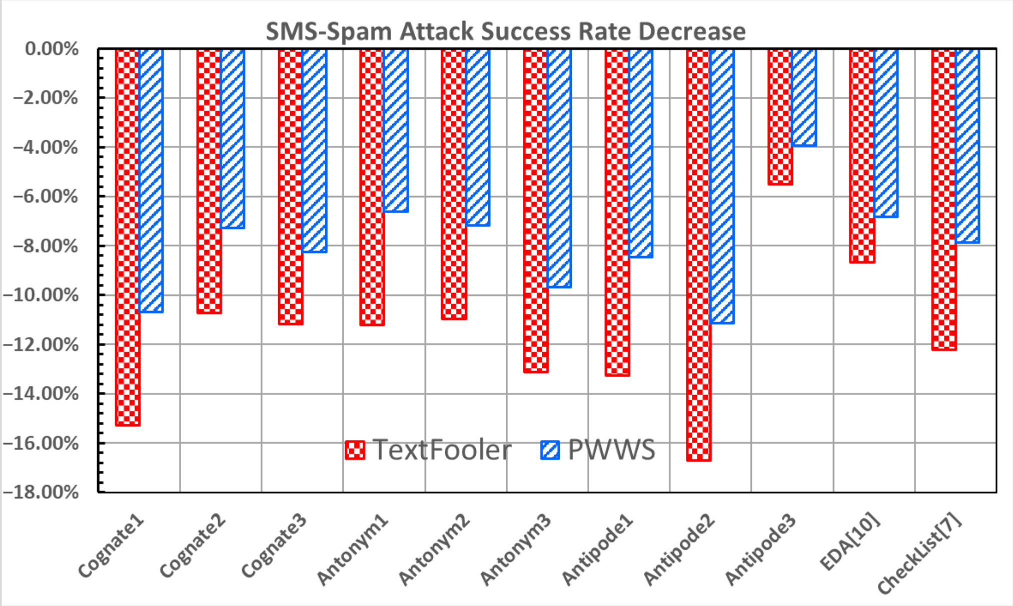

Figure 12. SMS-Spam Attack Success Rate Decrease.

Our proposed data augmentation methods, particularly the Antonym-based and Antipode-based variations, demonstrate superior performance in terms of the robustness improvement compared to the current methods EDA [10] and CheckList [7].

5. Additional Analysis

The Antipode3 models in Figure 1 exhibit the highest average performance improvement, with the Antonym2 models being the only exception that did not experience a decrease in the average accuracy. To further analyze these unusual results, we employed a confusion matrix to compare the original and three Antipode3 models for AG-News, the original model, three Antonym2 models, and the two lowest-performing models for TREC. This analysis allows us to examine the influence of our data augmentation methods on the model's performance. The accuracy and F1 scores of these models on AG-News and TREC can be found in Tables 2 and 3. Run *n* means the nth evaluation.

Table 2. Target Models on AG-News.

Models	Accuracy	F1-Score
Original Model	90.00%	89.80%
Antipode3 Run1	92.40%	92.27%
Antipode3 Run2	91.80%	91.62%
Antipode3 Run3	91.40%	91.23%

Table 3. Target Models on TREC.

Models	Accuracy	F1-Score
Original Model	97.00%	95.53%
Antonym2 Run1	97.40%	94.73%
Antonym2 Run2	96.80%	94.26%
Antonym2 Run3	97.00%	94.42%
Antonym3 Run1	95.60%	92.46%
Antipode3 Run3	95.80%	93.29%

5.1. AG-News

Figures 13 and 14 show the confusion matrices for the original model and Antipode3 Run1 (which has the best performance). The confusion matrices for the other models can be found in Figures A1 and A2.

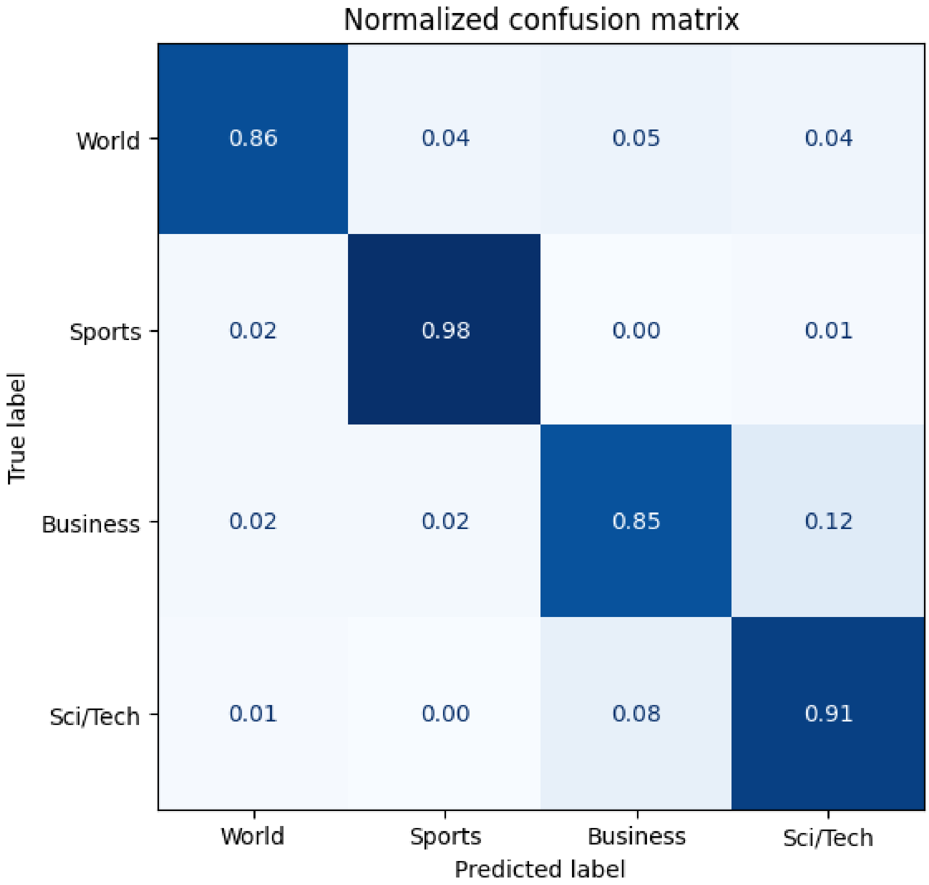

Figure 13. Original Model Confusion Matrix on AG-News.

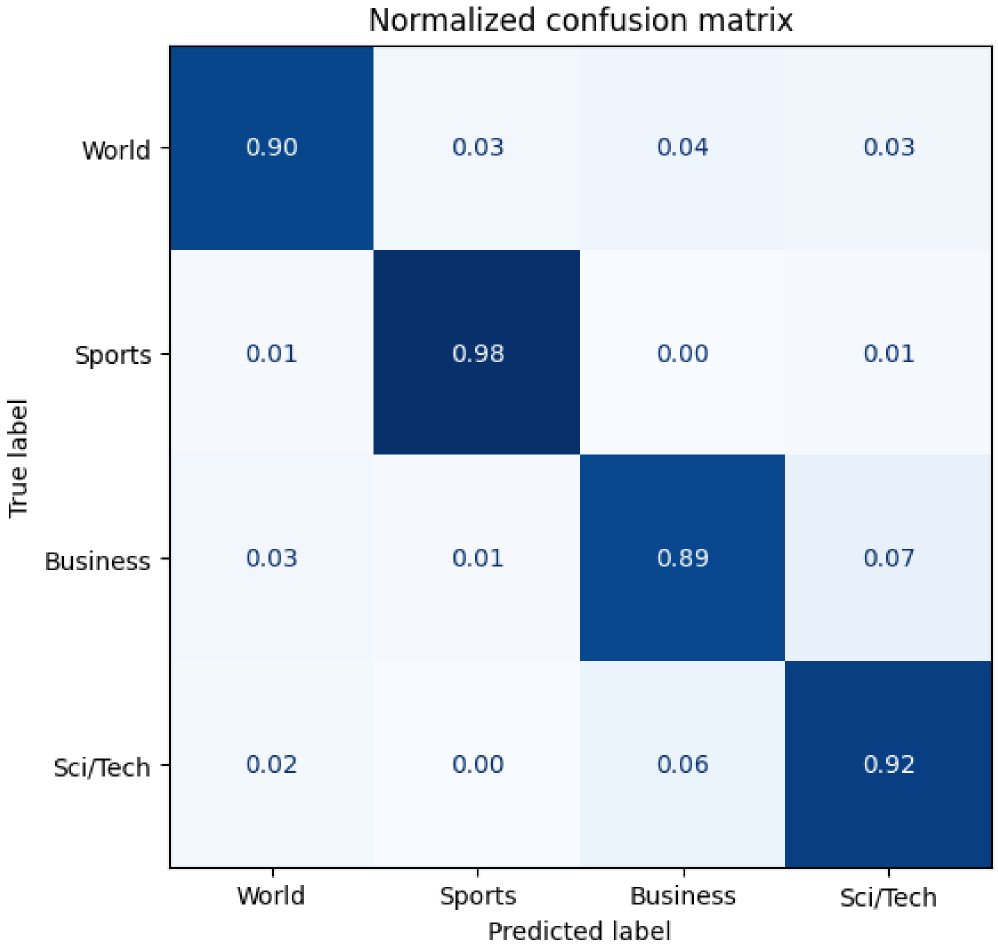

Figure 14. Antipode3 Run1 Confusion Matrix on AG-News.

When comparing the original model with the three Antipode3 models (Run1), the best model is the most effective, and the direction of improvement is consistent across all of them: they all demonstrate an increased accuracy in recognizing the World class, and there is less confusion between the Business and Sci/Tech classes. This improvement may be due to the expansion of vocabulary through data proliferation by Antipode3, which has reduced the attention given to words that are unrelated to the label.

5.2. TREC

Figures 15–17 show the confusion matrices for the original model, Antonym2 Run1 (which has the best performance), and Antonym3 Run1 (which has the worst performance). The confusion matrices for the other models can be found in Figures A3–A5.

Upon comparing Antipode1, Antipode2, and Antipode3, it was found that Antipode1 has the most diverse lexical variations, but Antipode2 is the most effective overall. This finding emphasizes that while an increased lexical diversity can be beneficial, it can also introduce noise into the data, potentially decreasing the model's performance and robustness. In some instances, our methods resulted in a reduction in the performance and robustness. Our additional analysis also indicates that the generated data may hinder normal learning, leading to a decreased performance. Therefore, it is essential to weigh the benefits of lexical diversity against the potential negative impacts of noise in future research. To further examine this trade-off and minimize adverse effects while preserving the benefits of increased lexical diversity, we plan to implement more constraints and conduct additional experiments. While our focus is not necessarily on developing the optimal model, we believe that incorporating our data augmentation methods with optimal parameters and task-specific pre-trained models has the potential to generate more accurate and robust models. While our methods may not be directly applied to specific tasks, such as sentiment analysis, they can be modified through label- and word-swap constraints to be more suitable for these tasks.

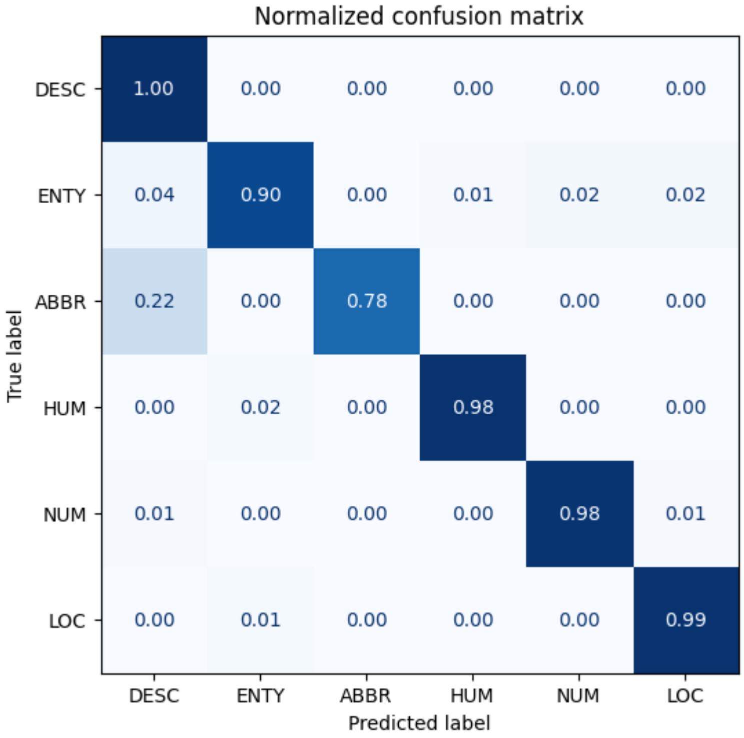

Figure 15. Original Model Confusion Matrix on TREC.

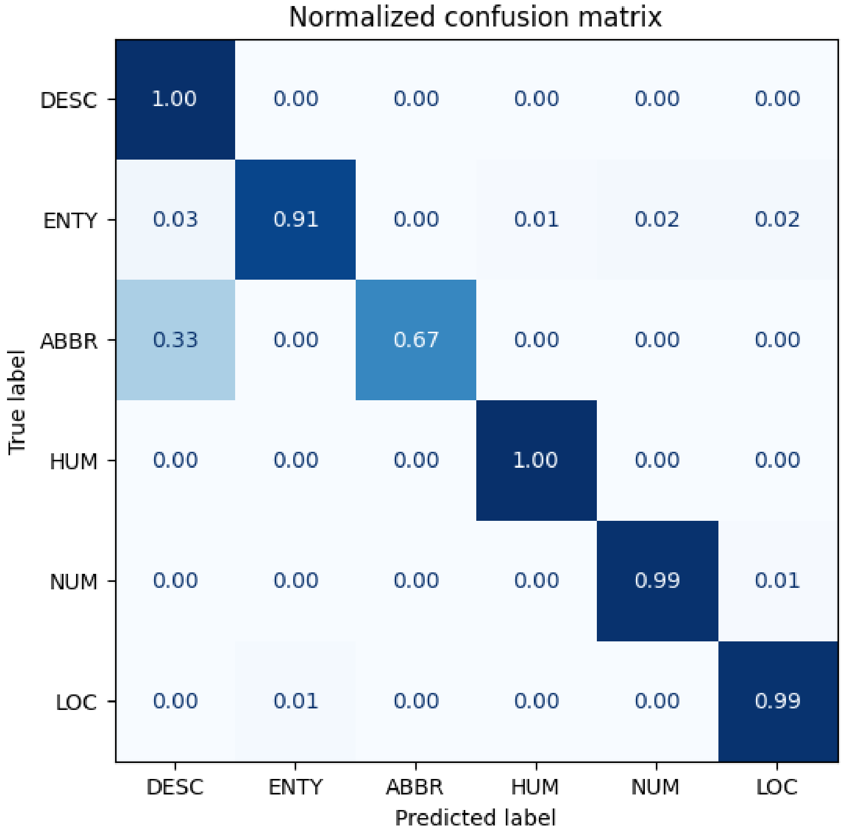

Figure 16. Antonym2 Run1 Confusion Matrix on TREC.

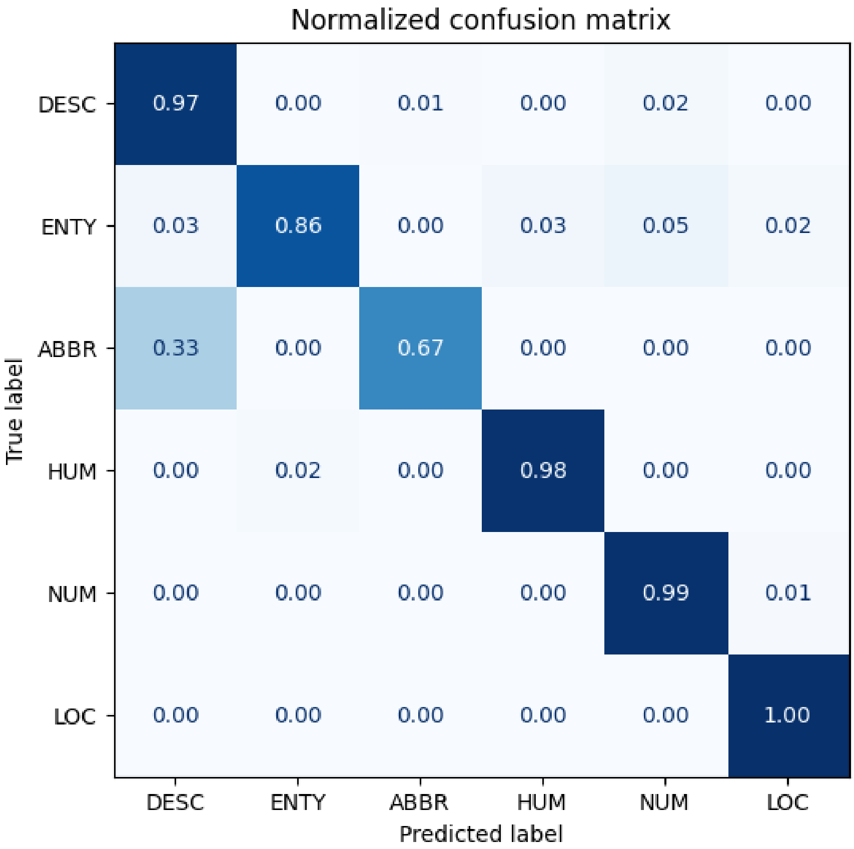

Figure 17. Antonym3 Run1 Confusion Matrix on TREC.

Our Antipode3 method demonstrates its superiority in the AG-News section, but it does not happen in this TREC section. Upon comparing the original model with the Antonym2 models, we see that the direction of improvement and decrease are consistent, with Antonym2 improving the overall accuracy but significantly increasing the confusion in recognizing ABBR as DESC. Both categories are used for identification, with the latter containing the former, which may explain the low original accuracy. When comparing the original model with worse models, we observe a similar decline in performance but without any improvement or even a further decrease. Our proposed methods may impede the learning of words related to the label but prevent the learning of unrelated words simultaneously. Our best model has more benefits than drawbacks, and other models are vice versa.

6. Discussion

We evaluated the effectiveness of our proposed data augmentation methods in terms of the performance and robustness. The results showed that Antipode-based methods are beneficial for improving the performance, with both Antonym-based and Antipode-based methods exhibiting an enhanced robustness. The Antipode-based method, in particular, demonstrated a superior performance and robustness enhancement compared to prior data augmentation methods that utilize similar lexical words to increase diversity.

To determine the significance of diverse semantics and lexical variations, we compared the results of our proposed Cognate-based, Antonym-based, and Antipode-based methods. The Cognate-based methods have a greater lexical diversity, while the Antonym-based methods have diverse semantics and the Antipode-based methods have diverse semantics and lexical variations. Our Antonym-based and Antipode-based methods demonstrated that diverse semantics are essential for improving the robustness. On the other hand, diverse semantics and lexical variations are crucial for a performance improvement. Our findings suggest that the use of antonyms and opposite sentiment words to create semantically and lexically diverse data may be helpful in text classification tasks in terms of both a performance and robustness improvement, especially when the original data are limited in its semantic and lexical variety (according to the results of SMS-Spam [24]). These results encourage a further exploration into data augmentation beyond similarity.

We will also explore the use of our methods in other natural language processing tasks requiring robustness, such as topic segmentation and authorship attribution, to see if they can be applied similarly to improve the performance and robustness.

7. Conclusions

In this study, we introduced three novel data augmentation methods for improving the robustness of text classification models. We conducted evaluation experiments on four text classification datasets and found that, in addition to an increased robustness, our augmented datasets improve the prediction models' performance. We also compared our methods to two existing data augmentation methods. We found that one of our proposed methods performs similarly in terms of a performance enhancement while demonstrating superior results in terms of a robustness enhancement. Our empirical results demonstrated the effectiveness of our proposed data augmentation methods.

Author Contributions: Conceptualization, H.T.; software, H.T.; validation, H.T.; formal analysis, H.T.; investigation, H.T.; data curation, H.T.; writing-original draft, H.T.; writing-review and supervision, S.K. and Y.M.; funding acquisition, Y.M.; project administration, Y.M.; All authors have read and agreed to the published version of the manuscript.

Funding: This research has been supported by KAKENHI (20K11830) Japan.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: https://github.com/tang7777777/MDPI_codes (accessed on 14 December 2022).

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. AG-News Performance and Robustness Results.

Models	Accuracy	F1-Score	Accuracy under Attack		Attack Success Rate	
			TextFooler	PWWS	TextFooler	PWWS
Original Model	90.00%	89.80%	17.00%	28.60%	81.11%	68.22%
Cognate1	90.67%	90.38%	16.53%	34.00%	81.76%	62.49%
Cognate2	90.33%	90.03%	17.33%	37.00%	80.81%	59.05%
Cognate3	90.27%	90.02%	18.27%	36.33%	79.76%	59.74%
Antonym1	90.40%	90.14%	15.40%	34.73%	82.96%	61.56%
Antonym2	90.73%	90.44%	19.33%	35.47%	78.68%	60.91%
Antonym3	90.73%	90.53%	17.33%	35.00%	80.90%	61.43%
Antipode1	90.07%	89.81%	12.27%	26.40%	86.38%	70.69%
Antipode2	90.47%	90.21%	19.80%	37.20%	78.11%	58.88%
Antipode3	91.87%	91.71%	14.87%	33.33%	83.81%	63.71%
EDA [10]	90.80%	90.61%	14.80%	33.60%	83.68%	62.98%
CheckList [7]	90.73%	90.51%	14.20%	29.80%	84.36%	67.15%

Table A2. TREC Performance and Robustness Results.

Models	Accuracy	F1-Score	Accuracy under Attack TextFooler	PWWS	Attack Success Rate TextFooler	PWWS
Original Model	97.00%	95.53%	30.20%	50.60%	68.87%	47.84%
Cognate1	97.00%	94.25%	28.87%	58.87%	70.24%	39.31%
Cognate2	96.80%	94.60%	31.07%	57.07%	67.90%	41.05%
Cognate3	96.67%	94.51%	29.47%	57.53%	69.52%	40.49%
Antonym1	96.73%	94.84%	31.27%	54.40%	67.68%	43.76%
Antonym2	97.07%	94.47%	31.87%	51.87%	67.17%	46.57%
Antonym3	96.07%	93.52%	29.67%	52.20%	69.12%	45.66%
Antipode1	96.40%	93.84%	31.00%	51.40%	67.84%	46.68%
Antipode2	96.80%	93.93%	29.47%	51.87%	69.56%	46.42%
Antipode3	96.27%	93.85%	30.33%	52.73%	68.49%	45.22%
EDA [10]	96.67%	94.49%	29.20%	52.27%	69.79%	45.92%
CheckList [7]	96.33%	93.79%	31.67%	50.80%	67.13%	47.26%

Table A3. SUBJ Performance and Robustness Results.

Models	Accuracy	F1-Score	Accuracy under Attack TextFooler	PWWS	Attack Success Rate TextFooler	PWWS
Original Model	95.80%	95.80%	23.20%	34.40%	75.78%	64.09%
Cognate1	95.93%	95.93%	20.60%	33.80%	78.52%	64.77%
Cognate2	95.93%	95.93%	20.60%	34.13%	78.53%	64.43%
Cognate3	95.87%	95.86%	21.47%	34.27%	77.61%	64.26%
Antonym1	96.20%	96.19%	19.73%	34.73%	79.48%	63.89%
Antonym2	95.67%	95.66%	24.33%	36.73%	74.58%	61.61%
Antonym3	95.67%	95.66%	20.67%	33.33%	78.39%	65.15%
Antipode1	95.87%	95.86%	24.20%	36.93%	74.76%	61.48%
Antipode2	95.87%	95.87%	21.07%	34.60%	78.03%	63.91%
Antipode3	95.40%	95.40%	21.67%	35.27%	77.29%	63.03%
EDA [10]	96.40%	96.40%	22.33%	34.27%	76.84%	64.46%
CheckList [7]	95.33%	95.33%	20.33%	32.40%	78.66%	66.00%

Table A4. SMS-Spam Performance and Robustness Results.

Models	Accuracy	F1-Score	Accuracy under Attack TextFooler	PWWS	Attack Success Rate TextFooler	PWWS
Original Model	99.00%	97.68%	72.40%	79.40%	26.87%	19.80%
Cognate1	99.33%	98.41%	87.80%	90.27%	11.61%	9.13%
Cognate2	99.40%	98.58%	83.33%	86.93%	16.16%	12.54%
Cognate3	99.33%	98.42%	83.73%	87.87%	15.71%	11.54%
Antonym1	99.20%	98.13%	83.67%	86.13%	15.66%	13.17%
Antonym2	99.27%	98.28%	83.47%	86.73%	15.92%	12.62%
Antonym3	99.27%	98.27%	85.60%	89.20%	13.77%	10.14%
Antipode1	99.33%	98.43%	85.80%	88.07%	13.62%	11.34%
Antipode2	99.47%	98.73%	89.33%	90.87%	10.19%	8.65%
Antipode3	99.13%	97.98%	77.93%	83.40%	21.38%	15.87%
EDA [10]	99.27%	98.28%	81.20%	86.40%	18.21%	12.96%
CheckList [7]	99.53%	98.90%	84.93%	87.67%	14.67%	11.92%

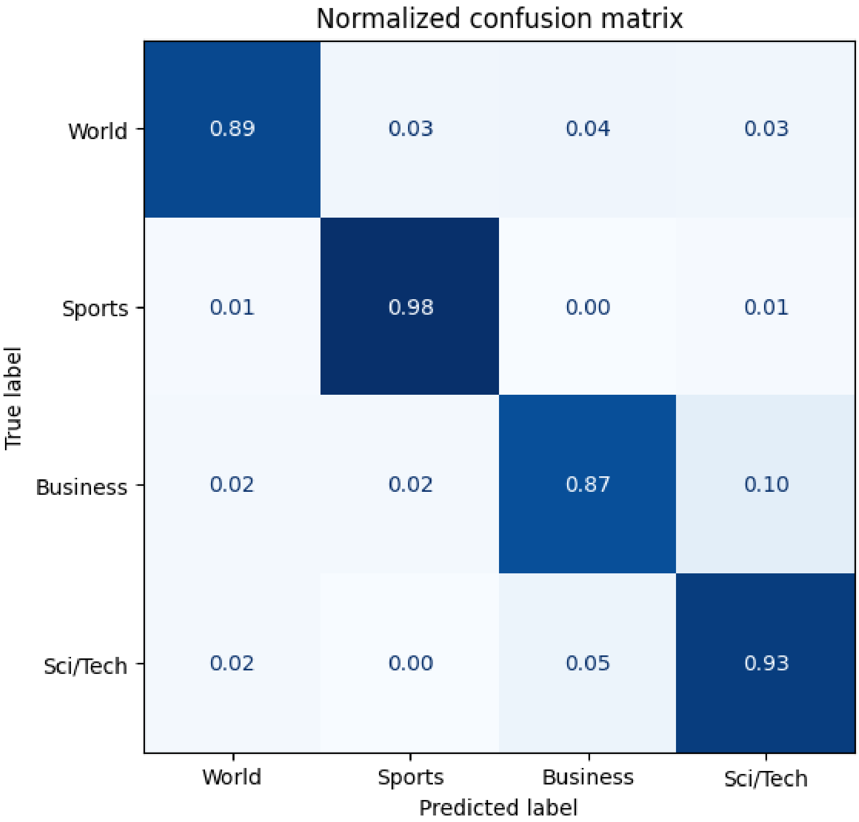

Figure A1. Antipode3 Run2 Confusion Matrix on AG-News.

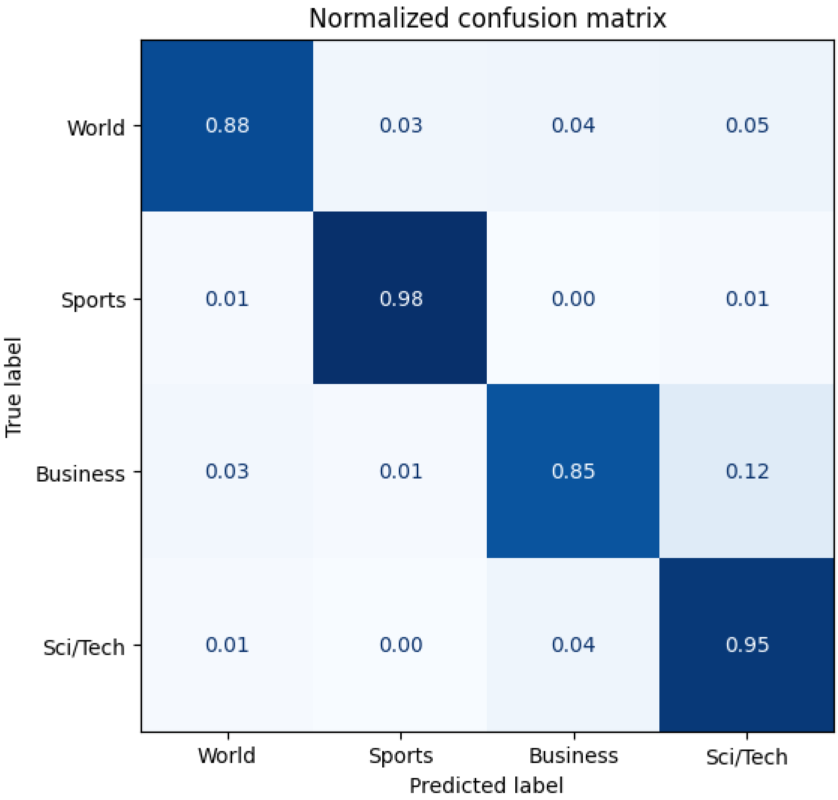

Figure A2. Antipode3 Run3 Confusion Matrix on AG-News.

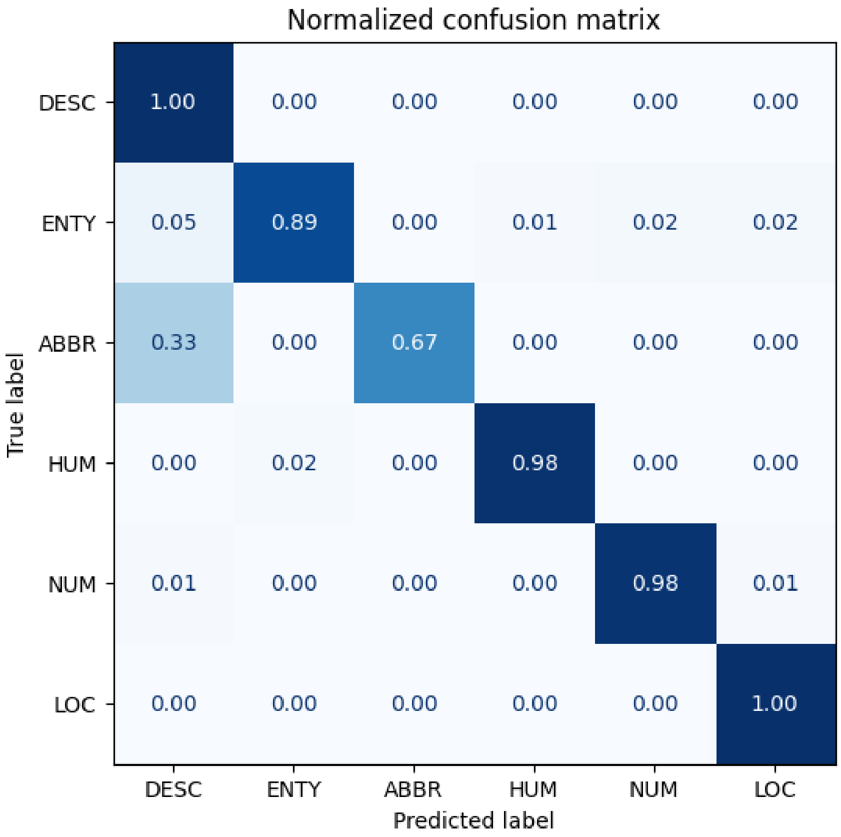

Figure A3. Antonym2 Run2 Confusion Matrix on TREC.

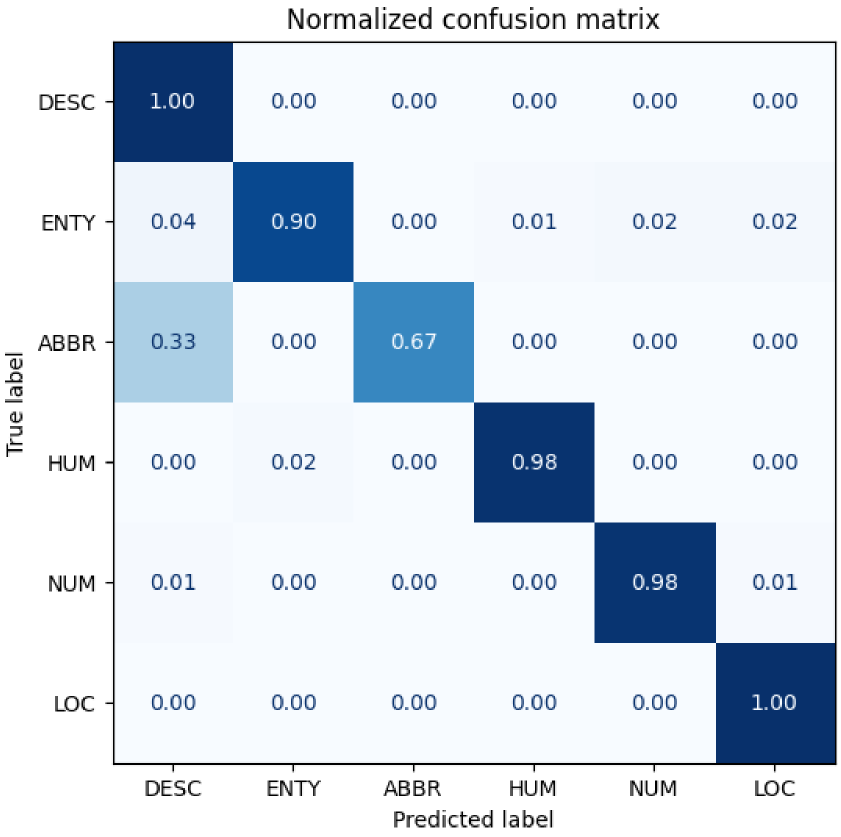

Figure A4. Antonym2 Run3 Confusion Matrix on TREC.

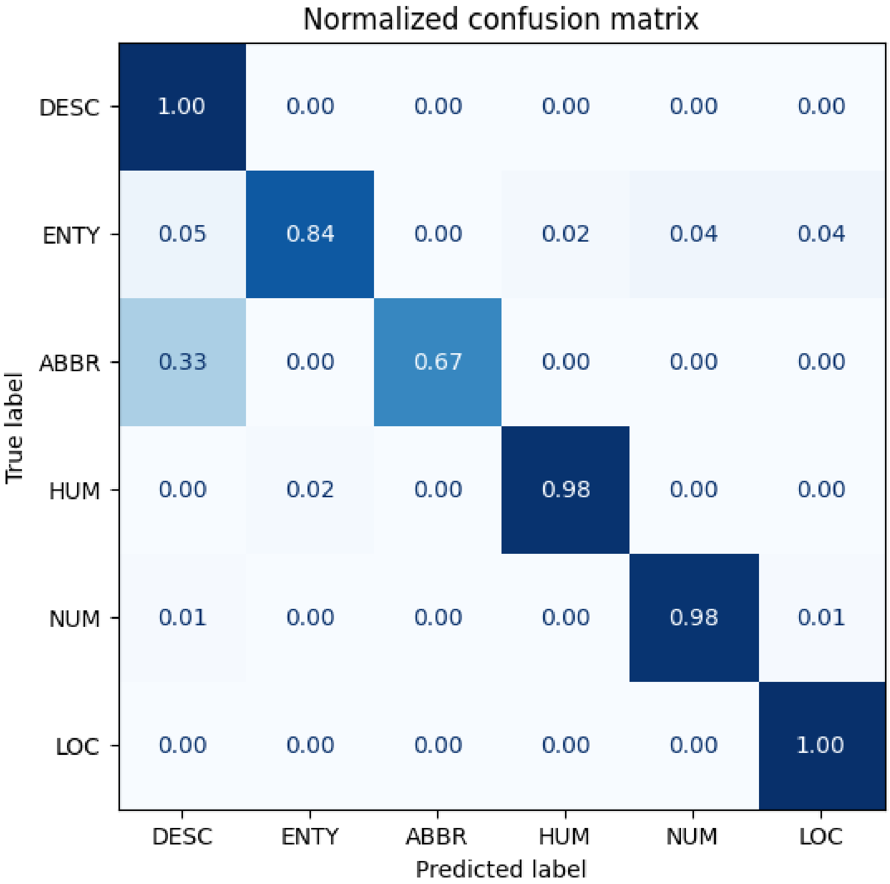

Figure A5. Antipode3 Run3 Confusion Matrix on TREC.

References

1. Faris, H.; Ala'M, A.-Z.; Heidari, A.A.; Aljarah, I.; Mafarja, M.; Hassonah, M.A.; Fujita, H. An intelligent system for spam detection and identification of the most relevant features based on evolutionary random weight networks. *Inf. Fusion* **2019**, *48*, 67–83. [CrossRef]
2. Daouadi, K.E.; Rebaï, R.Z.; Amous, I. Optimizing semantic deep forest for tweet topic classification. *Inf. Syst.* **2021**, *101*, 101801. [CrossRef]
3. Fan, F.; Feng, Y.; Zhao, D. Multi-grained Attention Network for Aspect-Level Sentiment Classification. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 3433–3442.
4. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota, 2–7 June 2019; pp. 4171–4186.
5. Sanh, V.; Debut, L.; Chaumond, J.; Wolf, T. DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv* **2019**, arXiv:1910.01108.
6. Jin, D.; Jin, Z.; Zhou, J.T.; Szolovits, P. Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 8018–8025.
7. Ribeiro, M.T.; Wu, T.; Guestrin, C.; Singh, S. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 4902–4912.
8. Li, D.; Zhang, Y.; Peng, H.; Chen, L.; Brockett, C.; Sun, M.-T.; Dolan, B. Contextualized Perturbation for Textual Adversarial Attack. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online, 6–11 June 2021; pp. 5053–5069.
9. Ren, S.; Deng, Y.; He, K.; Che, W. Generating Natural Language Adversarial Examples through Probability Weighted Word Saliency. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 1085–1097.
10. Wei, J.; Zou, K. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. 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), Hong Kong, China, 3–7 November 2019; pp. 6382–6388.

11. Karimi, A.; Rossi, L.; Prati, A. AEDA: An Easier Data Augmentation Technique for Text Classification. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2021, Punta Cana, Dominican Republic, 7–11 November 2021; pp. 2748–2754.
12. Liu, R.; Xu, G.; Jia, C.; Ma, W.; Wang, L.; Vosoughi, S. Data Boost: Text Data Augmentation Through Reinforcement Learning Guided Conditional Generation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; pp. 9031–9041.
13. Anaby-Tavor, A.; Carmeli, B.; Goldbraich, E.; Kantor, A.; Kour, G.; Shlomov, S.; Tepper, N.; Zwerdling, N. Do Not Have Enough Data? Deep Learning to the Rescue! In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 7383–7390.
14. Xie, Q.; Dai, Z.; Hovy, E.; Luong, T.; Le, Q. Unsupervised data augmentation for consistency training. In Proceedings of the 34th International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 6–12 December 2020; pp. 6256–6268.
15. Kobayashi, S. Contextual Augmentation: Data Augmentation by Words with Paradigmatic Relations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), New Orleans, LA, USA, 1–6 June 2018; pp. 452–457.
16. Şahin, G.G.; Steedman, M. Data Augmentation via Dependency Tree Morphing for Low-Resource Languages. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 5004–5009.
17. Niu, T.; Bansal, M. Automatically Learning Data Augmentation Policies for Dialogue Tasks. 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), Hong Kong, China, 3–7 November 2019; pp. 1317–1323.
18. de Arruda, H.F.; Costa, L.d.F.; Amancio, D.R. Topic segmentation via community detection in complex networks. *Chaos: Interdiscip. J. Nonlinear Sci.* **2016**, *26*, 063120. [CrossRef] [PubMed]
19. Machicao, J.; Corrêa Jr, E.A.; Miranda, G.H.; Amancio, D.R.; Bruno, O.M. Authorship attribution based on life-like network automata. *PLoS ONE* **2018**, *13*, e0193703. [CrossRef] [PubMed]
20. Zhang, X.; Zhao, J.; LeCun, Y. Character-level convolutional networks for text classification. In Proceedings of the 28th International Conference on Neural Information Processing Systems—Volume 1, Montreal, Quebec, Canada, 7–12 December 2015; pp. 649–657.
21. Li, X.; Roth, D. Learning Question Classifiers. In Proceedings of the 19th International Conference on Computational Linguistics, Taipei, Taiwan, 24 August–1 September 2002.
22. Hovy, E.; Gerber, L.; Hermjakob, U.; Lin, C.-Y.; Ravichandran, D. Toward semantics-based answer pinpointing. In Proceedings of the First International Conference on Human Language Technology Research, San Diego, CA, USA, 18–21 March 2001; pp. 1–7.
23. Conneau, A.; Kiela, D. SentEval: An Evaluation Toolkit for Universal Sentence Representations. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 7–12 May 2018.
24. Almeida, T.A.; Hidalgo, J.M.G.; Yamakami, A. Contributions to the study of SMS spam filtering: New collection and results. In Proceedings of the 11th ACM symposium on Document engineering, Mountain View, CA, USA, 19–22 September 2011; pp. 259–262.
25. Miller, G.A. WordNet: A lexical database for English. *Commun. ACM* **1995**, *38*, 39–41. [CrossRef]
26. Cambria, E.; Li, Y.; Xing, F.Z.; Poria, S.; Kwok, K. SenticNet 6: Ensemble application of symbolic and sub-symbolic AI for sentiment analysis. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management, Online, 19–23 October 2020; pp. 105–114.
27. Morris, J.X.; Lifland, E.; Yoo, J.Y.; Grigsby, J.; Jin, D.; Qi, Y. TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Online, 16–20 November 2020; pp. 119–126.
28. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. Transformers: State-of-the-Art Natural Language Processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Online, 16–20 November 2020; pp. 38–45.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.