Selective Word Substitution for Contextualized Data Augmentation

Kyriaki Pantelidou, Despoina Chatzakou, Theodora Tsikrika
Stefanos Vrochidis, Ioannis Kompatsiaris

Information Technologies Institute, Centre for Research and Technology Hellas
{kpantelidou,dchatzakou,theodora.tsikrika,stefanos,ikom}@iti.gr

Abstract. The often observed unavailability of large amounts of training data typically required by deep learning models to perform well in the context of NLP tasks has given rise to the exploration of data augmentation techniques. Originally, such techniques mainly focused on rule-based methods (e.g. random insertion/deletion of words) or synonym replacement with the help of lexicons. More recently, model-based techniques which involve the use of non-contextual (e.g. Word2Vec, GloVe) or contextual (e.g. BERT) embeddings seem to be gaining ground as a more effective way of word replacement. For BERT, in particular, which has been employed successfully in various NLP tasks, data augmentation is typically performed by applying a masking approach where an arbitrary number of word positions is selected to replace words with others of the same meaning. Considering that the words selected for substitution are bound to affect the final outcome, this work examines different ways of selecting the words to be replaced by emphasizing different parts of a sentence, namely specific parts of speech or words that carry more sentiment information. Our goal is to study the effect of selecting the words to be substituted during data augmentation on the final performance of a classification model. Evaluation experiments performed for binary classification tasks on two benchmark datasets indicate improvements in the effectiveness against state-of-the-art baselines.

Keywords: Contextual data augmentation, word substitution, binary classification

1 Introduction

Due to the large amount of text data available in the wild, natural language processing (NLP) techniques have been developed to automatically comprehend and represent human language. To this end, machine learning models are often built, with deep neural networks being among the most prominent, due to their increased effectiveness in a multitude of tasks. However, the performance of deep learning models depends significantly on the size of the ground truth dataset [3]; creating though a large ground truth dataset is a rather expensive and time consuming process. To overcome the lack of large annotated datasets, data augmentation methods have emerged to artificially expand the size of such datasets,

primarily in the areas of computer vision [15, 6] and speech [9], with recent works focusing also on textual data [20, 10, 18]. For example, multi-granularity text-oriented data augmentation was recently proposed, including word-, phrase-, and sentence-level data augmentation [18]. Moreover, the "Easy Data Augmentation" (EDA) method suggested four operations towards data augmentation: synonym replacement, random insertion, random swap, and random deletion [20].

As an alternative, contextualized data augmentation techniques have arisen due to their ability to better capture the contextual relations between words in a sentence, thus resulting in a more effective word suggestion for substitution. For instance, the so-called Contextual Augmentation follows a contextual prediction approach to suggest words that have paradigmatic relations with the original words [10]. More recent studies on data augmentation utilize BERT (Bidirectional Encoder Representations from Transformers) [7], a pre-trained deep bidirectional representation that jointly conditions on both left and right context in all layers. For example, the proposed "masked language model" (MLM) randomly masks a certain percentage of input tokens to finally predict such masked words based on their context for the formation of new sentences [4]. A complementary approach is the so-called "Conditional BERT Contextual Augmentation", a variant of the BERT-based model aimed at predicting label-compatible words based on both their context and sentence label [21]; however, its close tie with the BERT architecture somewhat prevents its direct use in other pre-trained language models. As a countermeasure, a solution that prepends the class label to text sequences has been suggested, resulting in a simple but yet effective way to condition the pre-trained models for data augmentation [11]. Finally, more recently, the "Generative pre-training" (GPT-2) model has been used as a basis for artificially generating new labeled data [1].

Thus far, contextualized BERT-based methods have only considered random selections of words for replacement in order to generate new synthetic sentences. As intuitively expected, the words selected for replacement can play a decisive role in the quality of the new generated sentences that are subsequently used as input for training a neural network model. In this regard, this work proposes more targeted word replacements, by paying attention to specific and potentially more significant and/or informative parts of a sequence. In particular, this work proposes that specific part-of-speech tags, such as adjectives and adverbs that can be seen as modifiers of other words (e.g. nouns, pronouns and verbs), are considered for replacement, as well as words that carry sentiment information. Our intuition is that focusing only on such words for substitution will lead to new sentences that will provide more variability in a deep learning model for better adaptation to different ways of expressing the same thing; methods adopted so far in most of cases lead to the substitution of words that are not always sufficiently distinctive and therefore may not result in substantially different sentences. Evaluation is performed for binary classification tasks on two well-established datasets, the SST-2 [16] and the SUBJ [13], which comprise of short sentences characterized as either positive/negative or subjective/objective,

respectively. Overall, the experimental results underscore the contribution and positive affect of proper selection of words for replacement in data augmentation.

2 Methodology

This section outlines: (i) the methods considered for word substitution (both the baseline and the ones proposed by this work); (ii) the contextualized data augmentation method applied, following a label-compatible or a label-independent replacement approach; and finally (iii) the deep neural network model used to evaluate the proposed approaches for word selection and substitution.

2.1 Substitution of Words: target masked words

As already mentioned, in this work, BERT-based models are utilized for data augmentation. Before feeding word sequences into BERT-based models, part of words in a sequence is replaced with a [MASK] token. The model then attempts to predict the original value of the mask words, based on the context provided by the rest, non-masked, words in the sequence, thus resulting to the creation of new sentences with the same meaning but different words. So, the first step involves the selection of words to be masked. To this end, next we describe the baseline (i.e. Random) and the three proposed approaches (i.e. Word modifiers (all), Word modifiers (15%), and Sentiment-related words) for word selection and substitution thereafter.

- **Random:** we randomly mask at maximum the 15% of all word tokens in a sequence. We iterate over all tokens in a sequence and when the i-th token is chosen, we replace it with (i) the [MASK] token 80% of the time and (ii) the unchanged i-th token 20% of the time. We repeat this process until the 15% of the tokens are masked or when all the tokens have been checked. This typical approach for contextual data augmentation [4] is our baseline.
- **Word modifiers (all):** we obtain the part-of-speech (POS) tag of each token in a sequence using the spaCy library [17]. Then, we choose to mask all tokens characterized as ADJ (adjectives) and ADV (adverbs), since these function as word modifiers for nouns and verbs, respectively, and could therefore influence and have a greater impact on the actual meaning of a sentence.
- **Word modifiers (15%):** we follow the same process as before, with the difference being the selection of a maximum of 15% of ADJ and ADV tokens to be masked. Similar to the baseline method, 80% of the times the token is marked with [MASK] and 20% it remains unchanged.
- **Sentiment-related words:** we utilize SentiWordNet [5], a lexical resource in which each word is associated with three numerical scores Obj, Pos, and Neg, describing how objective, positive, and negative the word is. The sum of these scores equals to 1.0. We choose to mask all the words that have a positive or negative score > 0.0, excluding this way the more "neutral" terms that have no-influence on the sentiment of a sentence.

Hence, at the end of each of the aforementioned words selection processes, we obtain a list with the target masked words to be predicted by the BERT-based models.

2.2 Contextualized Data Augmentation

For data augmentation, we considered both BERT and conditional BERT.

BERT-based approach. For the prediction of the target masked words, first we proceed with BERT [4]. In particular, the "bert-base-uncased" model [2] is utilised, i.e. a pretrained model on English language using a masked language modeling (MLM) objective.

Conditional BERT-based approach (CBERT). The conditional BERT-based contextual augmentation [21] considers the label of the original sequence for artificially generating new labeled data. CBERT shares the same model architecture with the original BERT. The main differences lay on the input representation and training procedure. The input embeddings of BERT are the sum of the token embeddings, the segmentation embeddings, and the position embeddings. However, these embeddings have no connection to the actual annotated labels of a sentence, thus the predicted word is not always compatible with the annotated label. To build a conditional MLM, CBERT finetunes on the pre-trained BERT (i.e. "bert-base-uncased") by altering the segmentation embeddings to label embeddings, which are learned corresponding to their annotated labels on labeled datasets. In the end, the model is expected to be able to predict words in masked position by considering both the context and the label.

2.3 Deep Neural Network Model

To evaluate the proposed alternatives for word substitution in a contextualized data augmentation setup, we focus on a binary classification task. In this respect, and in the same direction with similar works [10, 20, 21], a Convolutional Neural Network (CNN) based architecture is built given that CNNs have proven to be invaluable for NLP tasks due to their ability to capture salient information from n-gram word combinations.

Text input. Before feeding any text to the network, a set of preprocessing steps takes place to reduce noise. Specifically, URLs, digits, and single character words are removed, as well as punctuation and special characters. In addition, as neural networks are trained in mini-batches, each batch in the sample should have the same sequence length; we set the sequence length to 100.

Embedding layer. The first layer of the neural network architecture is the embedding layer, which maps each word to a high-dimensional layer. We opted for pre-trained word embeddings from GloVe [14] of 100 dimensions.

Neural Network layer. The CNN architecture is as follows: one 1D convolutional layer of 50 filters and kernel size 3, 1D average pool layer, and one dense layer of 20 hidden units. To avoid over-fitting, we use a spatial dropout of

$p = 0.6$. This architecture was chosen as it leads to fair performance across the two datasets considered.

Classification layer. Since in our case the objective is to classify texts into two classes, *sigmoid* is used as activation function.

3 Experimental Setup and Results

3.1 Datasets

We conducted experiments on two benchmark datasets: (i) the **SST-2**: Stanford Sentiment Treebank [16], which includes one-sentence movie reviews labeled positive or negative. The dataset contains 8,741 samples and is split in (train, test, validation) sets with size (6,228, 1,821, 692), respectively; and (ii) the **SUBJ**: Subjectivity Dataset [13] which includes 5,000 subjective and 5,000 objective processed sentences of movie reviews and it is split in (9,000, 900, 100) samples for (train, test, validation), respectively. After data augmentation, we doubled the size of the training sets, while keeping the test and validation sets intact.

3.2 Experimental Setup

For our implementation, we use Keras [8] with TensorFlow [19]. We run the experiments on a server with one GeForce RTX 2080 GPU of 12GB memory.

In terms of training, we use binary cross-entropy as loss function. As for the learning rate, which is a very important hyper-parameter when it comes to training a neural network, a scheduler-based approach is followed. Specifically, cosine annealing is used which initiates with a large learning rate that decreases relatively quickly to a minimum value before increasing rapidly again. For such an implementation, AdamW [12] is used. Finally, a maximum of 150 epochs is allowed, while also the validation set is used to perform early stopping. Training is interrupted if the validation loss does not drop in 3 consecutive epochs.

Experimentation phases. Overall, we consider two experimentation phases: (i) BERT-based contextual augmentation, and (ii) Conditional BERT-based contextual augmentation. For both phases, the four methods for words substitution presented in Section 2.1 are examined. For evaluation purposes, standard metrics are considered, namely accuracy (Acc), precision (Prec), recall (Rec), weighted area under the ROC curve (AUC), and the loss/accuracy curves.

3.3 Experimental Results

Table 1 and Table 2 present the results obtained with the classification model described in Section 2.3 before and after data augmentation is applied, for the SST-2 and SUBJ datasets, respectively. In all cases, the classification performance is better when an augmented dataset is used compared to when the original dataset is used, which highlights the overall usefulness of data augmentation.

BERT-based augmentation: In the first experimentation phase, we observe that the best performance in both datasets is achieved when specific (rather than random) words are selected for replacement.

For the SST-2 dataset, the best performance is obtained when only the 15% of the word modifiers is selected for replacement (acc: 82.12%, prec: 82.46%, rec: 82.13%, AUC: 90.76%); the differences between the Modifiers (15%) and the baseline are statistically significant ($p < 0.05$) for all evaluation metrics.

By analyzing the content of the augmented sentences produced by the different word substitution methods, we observe that the Random method replaces in most cases general words (such as pronouns and nouns) that do not affect the actual meaning and sentiment of a sentence. On the contrary, the Modifiers (all) and Sentiment-related words methods in many cases suggest words for replacement that change the overall sentiment of the sentence, resulting in lower performance compared to the random-based replacement approach (but still better than the non-augmented case). On the other hand, manual inspection of the sentences generated by the Modifiers (15%) method indicates that the number of cases where the sentiment changes is smaller compared to the other two proposed methods, while at the same time more informative words (of the same meaning) are replaced, thus resulting in the best overall performance.

Overall, it appears that the addition of sentences that have the same sentiment with the original ones and at the same time differ somehow in the way of expressing the same opinion and attitude in relation to an event (e.g. opinion about a movie as in our case) can probably make the model more capable of locating the correctly expressed sentiment on new data as it has learned to better interpret different ways of expressing the same or similar opinions and notions.

For the SUBJ dataset, we observe that in relation to accuracy, precision, and recall the selection of sentiment words for substitution allows for the generation of an augmented dataset that leads to a better performance ($p < 0.05$ when compared to the Random baseline). A slightly better performance in terms of AUC is achieved by the Modifiers (15%); compared though to the rest of the proposed methods their differences are not statistically significant ($p > 0.05$).

Contrary to the SST-2 dataset, in the SUBJ dataset, we observe that the selection of sentiment words for substitution leads to a better performance. As the SUBJ dataset is oriented towards the characterization of texts as either objective or subjective, it is quite expected that in subjective sentences the expression of sentiments/emotions will be more intense compared to the objective ones. Therefore, in this case, the presence of sentiment words alone in a sentence can be an important indication of the subsequent characterization of a text as subjective or objective, even without the consideration of labels during augmentation, which is examined next.

Conditional BERT-based augmentation: The aim in this experimental setup is to assess whether and how the inclusion of the label during augmentation will affect the overall performance in terms of the different word substitution approaches examined in this work. The results in both Table 1 and Table 2 indicate that when we focus on specific words for replacement we achieve better perfor-

Table 1. Classification results for the SST-2 dataset

	Before data augmentation			
	Acc	**Prec**	**Rec**	**AUC**
	80.79	81.19	80.80	89.45

	After data augmentation							
	BERT				**Conditional Bert**			
	Acc	**Prec**	**Rec**	**AUC**	**Acc**	**Prec**	**Rec**	**AUC**
Random (baseline)	81.85	81.92	81.85	90.50	81.58	81.95	81.59	90.28
Modifiers (all)	81.75	82.02	81.76	90.42	82.58	82.73	82.58	90.62
Modifiers (15%)	**82.12**	**82.46**	**82.13**	**90.76**	81.26	81.73	81.27	90.30
Sentiment-related words	81.61	81.84	81.62	90.34	**82.61**	**82.88**	**82.62**	**90.91**

Table 2. Classification results for the SUBJ dataset

	Before data augmentation			
	Acc	**Prec**	**Rec**	**AUC**
	90.53	90.54	90.53	96.38

	After data augmentation							
	BERT				**Conditional Bert**			
	Acc	**Prec**	**Rec**	**AUC**	**Acc**	**Prec**	**Rec**	**AUC**
Random (baseline)	91.18	91.19	91.18	96.49	90.78	90.80	90.78	96.81
Modifiers (all)	91.37	91.38	91.37	96.89	91.00	91.00	91.00	**96.83**
Modifiers (15%)	91.24	91.24	91.24	**96.91**	91.06	91.06	91.06	**96.83**
Sentiment-related words	**91.56**	**91.57**	**91.56**	96.88	**91.14**	**91.15**	**91.14**	**96.83**

mance in most cases compared to the baseline (random replacement). The best performance is obtained when only the sentiment words are considered for substitution. In Table 1, the Sentiment-related words method achieves acc: 82.61%, prec: 82.88%, rec: 82.62%, AUC: 90.91%, with the differences to the baseline being statistically significant ($p<0.05$) for all the evaluation metrics. In Table 2, the Sentiment-related words method achieves acc: 91.14%, prec: 91.15%, rec: 91.14%, AUC: 96.83%, with with the differences to the baseline being statistically significant ($p<0.05$) for all evaluation metrics except AUC. These results highlight the importance of focusing on sentiment words for substitution (probably due to such words conveying more useful information to the classification model), considering also the label of the original sequence.

4 Conclusions and Future Work

In this work we studied how the performance of a neural network model is affected when a more focused (rather than random) word substitution approach is adopted during contextualized data augmentation. The focus was on specific POS tags, as well as on words that reflect sentiments, following either a label-compatible or a label-independent replacement approach. Overall, the results indicate that a careful selection of words for substitution is beneficial. In the future, we intend to conduct a similar study by using alternative and more recent contextualized data augmentation methods, such as XLNet [22], which are able to capture the semantic relations between the masked tokens of a sequence.

Acknowledgements

This research is part of projects that have received funding from the European Union's H2020 research and innovation programme under CREST (GA No 833464), CONNEXIONs (GA No 786731), and STARLIGHT (GA No 101021797).

References

1. 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: AAAI Conference on Artificial Intelligence. vol. 34, pp. 7383–7390 (2020)
2. BERT base model: https://huggingface.co/bert-base-uncased, Accessed: 2021
3. Brownlee, J.: https://bit.ly/2SB0n5G (2019)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: NAACL (2019)
5. Esuli, A., Sebastiani, F.: Sentiwordnet: A publicly available lexical resource for opinion mining. In: Proceedings of the Fifth International Conference on Language Resources and Evaluation (LRECâĂŹ06) (2006)
6. Frid-Adar, M., Klang, E., Amitai, M., Goldberger, J., Greenspan, H.: Synthetic data augmentation using gan for improved liver lesion classification. In: 15th international symposium on biomedical imaging (ISBI'18). pp. 289–293. IEEE (2018)
7. Kenton, J.D.M.W.C., Toutanova, L.K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT. pp. 4171–4186 (2019)
8. Keras: https://keras.io/ (2020)
9. Ko, T., Peddinti, V., Povey, D., Khudanpur, S.: Audio augmentation for speech recognition. In: Sixteenth Annual Conference of the International Speech Communication Association (2015)
10. Kobayashi, S.: Contextual augmentation: Data augmentation by words with paradigmatic relations (2018)
11. Kumar, V., Choudhary, A., Cho, E.: Data augmentation using pre-trained transformer models. In: Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems. pp. 18–26. ACL (Dec 2020)
12. OverLordGoldDragon: Keras adamw. GitHub. Note: https://github.com/OverLordGoldDragon/keras-adamw/ (2019)
13. Pang, B., Lee, L.: A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. CoRR **cs.CL/0409058** (2004)
14. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)
15. Simard, P.Y., LeCun, Y.A., Denker, J.S., Victorri, B.: Transformation invariance in pattern recognitionâĂŤtangent distance and tangent propagation. In: Neural networks: tricks of the trade, pp. 239–274. Springer (1998)
16. Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C.D., Ng, A., Potts, C.: Recursive deep models for semantic compositionality over a sentiment treebank. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. pp. 1631–1642. Association for Computational Linguistics, Seattle, Washington, USA (Oct 2013)
17. SpaCy: https://spacy.io/, Accessed: 2021

18. Sun, X., He, J.: A novel approach to generate a large scale of supervised data for short text sentiment analysis. Multimedia Tools and Applications **79**(9), 5439–5459 (2020)
19. TensorFlow: https://www.tensorflow.org/, Accessed: 2021
20. 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). pp. 6383–6389. Association for Computational Linguistics, Hong Kong, China (Nov 2019)
21. Wu, X., Lv, S., Zang, L., Han, J., Hu, S.: Conditional bert contextual augmentation. In: Computational Science – ICCS 2019. pp. 84–95. Springer International Publishing, Cham (2019)
22. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R.R., Le, Q.V.: Xlnet: Generalized autoregressive pretraining for language understanding. In: Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc. (2019)