Word Substitution with Masked Language Models as Data Augmentation for Sentiment Analysis

Larisa Kolesnichenko
University of Oslo
larkkinn@gmail.com

Erik Velldal
University of Oslo
erikve@ifi.uio.no

Lilja Øvrelid
University of Oslo
liljao@ifi.uio.no

Abstract

This paper explores the use of masked language modeling (MLM) for data augmentation (DA), targeting structured sentiment analysis (SSA) for Norwegian based on a dataset of annotated reviews. Considering the limited resources for Norwegian language and the complexity of the annotation task, the aim is to investigate whether this approach to data augmentation can help boost the performance. We report on experiments with substituting words both inside and outside of sentiment annotations, and we also present an error analysis, discussing some of the potential pitfalls of using MLM-based DA for SSA, and suggest directions for future work.

1 Introduction

One important challenge in sentiment analysis, like for most areas of NLP approached as supervised learning tasks, is that of limited availability of labeled training data. As annotation is typically a manual process requiring human experts – thereby incurring a high cost in terms of time, effort, and money – the creation of labeled training data represents a major bottleneck, especially for smaller languages like Norwegian. At the same time, we know that the amount of training examples is the most important driver for increasing model performance. This paper reports on preliminary results with using a pre-trained masked language model (MLM) for data augmentation (DA), applying the MLM to generate alternative substitutions for different words in the training data. More specifically we investigate the role of MLM-based DA for the task of so-called *structured sentiment analysis* (SSA), i.e. predicting not just the positive/negative polarity of a text, but also the spans of polar expressions, targets and holders (Barnes et al., 2022).

We start by briefly discussing related work, before presenting the resources used in our experiments. We then discuss the details of the data augmentation strategy, before presenting our experimental results, including an error analysis.

2 Related work

Previous work on data augmentation in NLP can be viewed as either augmenting the feature space (e.g. interpolation techniques, addition of noise) versus augmentation of the (input) data space, where augmentation can take place at character, word, phrase or document-level (Bayer et al., 2022). While rule-based approaches to data augmentation have been studied, a majority of current DA-research within NLP falls within a model-based approach to augmentation of the input data (Feng et al., 2021). Examples of these techniques range from DA through back-translation (Sennrich et al., 2016) via seq2seq-based generation of paraphrases (Kumar et al., 2020) to the current widespread use of pre-trained LMs to generate text augmentations (Kobayashi, 2018; Yang et al., 2020). The latter approach is also the one adopted in this paper. Within the field of sentiment analysis, a specific challenge has been the issue of label preservation (Bayer et al., 2022), i.e. avoiding a switch of polarity following augmentation. Both lexicon-based (Hu et al., 2019) and embedding-based (Feng et al., 2019) methods for substitution, or combinations of the two, have been proposed to address this challenge.

The previous work that is perhaps closest to that of the current paper is that of Chen et al. (2022), both in terms of DA approach and application task. In the context of the 2022 SemEval Shared Task (task 10) on structured sentiment analysis (Barnes et al., 2022), Chen et al. (2022) augment the training datasets by masking out each token in turn and generating replacements with XLM-RoBERTa-large, only accepting the top 5 most

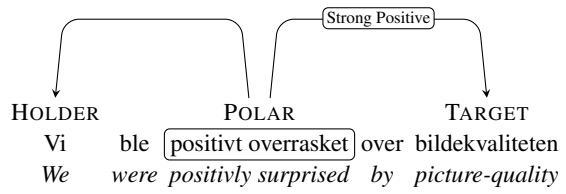

Figure 1: Example annotation from NoReC*fine*

confident word predictions. Chen et al. (2022) augments sentiment-bearing sentences only, but exclude replacements for tokens that are within the span of polar expressions.

3 Sentiment dataset and models

The Norwegian Review Corpus – NoReC (Velldal et al., 2018) – is a corpus of professional reviews gathered from multiple Norwegian news sources and spanning a wide range of different domains (books, music, movies, games, restaurants, various consumer goods, and more). A subset of this corpus, NoReC*fine* (Øvrelid et al., 2020), comprises 11 437 sentences that have been annotated for so-called structural or fine-grained sentiment information. The annotations indicate the span of polar expressions, their corresponding holder- and target expressions, and the associated polarity (positive/negative) and intensity (on a three-point scale), see Figure 1 for an example. Our experiments are performed on the pre-defined splits of the NoReC*fine* dataset.

All our experiments are based on the classification architecture described by Samuel et al. (2022), which adapts the graph-based semantic parser PERIN by Samuel and Straka (2020) for the task of structured sentiment analysis, achieving state-of-art results by directly predicting sentiment graphs from text. We will refer to this architecture as SSA-PERIN.[1]

A Norwegian instance of the BERT (base, cased) architecture (Devlin et al., 2019) dubbed NorBERT2 and released by the NorLM initiative (Kutuzov et al., 2021) is used both for training our SSA-PERIN models and also for generating substitutions in our MLM-DA experiments. It features a 50 000 WordPiece vocabulary and was trained using Whole Word Masking on the public part of the Norwegian Colossal Corpus (NCC) (Kummervold et al., 2022) and the Norwegian part of the mC4 corpus (Raffel et al., 2019), comprising a total of ≈ 15 billion tokens.

[1]We use the training configuration described in Samuel et al. (2022): AdamW optimizer (Loshchilov and Hutter, 2019) with the learning rate linearly warmed-up for the first 10% of the training span, and then decayed with a cosine schedule. We use the labeled edge-encoding and disabled character embeddings.

4 Data augmentation strategy

Our augmentation approach comprises three steps: (1) iterate through the sample text we want to augment, replacing each word in turn with a placeholder token; (2) prompt the LM to generate a replacement for the placeholder; and (3) if the probability for the prediction is above a given threshold p, add the sentence with the generated replacement to the augmented training set.

Our DA approach is similar to that of Chen et al. (2022), but differs in that we exhaustively consider replacements for all tokens in all sentences in the training data (not just the top 5 most probable replacements for sentiment-bearing sentences only), and we also test the effect of allowing replacements inside spans annotated as polar expressions. For the latter, we experiment with constraining the candidates by only allowing replacements that do not represent a switch of polarity with respect to the sentiment lexicon NorSentLex[2] (Barnes et al., 2019). In addition, we present separate experimental results for replacing tokens only *inside* or only *outside* annotation spans (target, holder, and polar expressions), and both, while also testing different values for the confidence threshold.

5 Experimental results and discussion

Below we first describe the tested model configurations and evaluations measures, before discussing the results and presenting an error analysis, including suggestions for future work.

Description of configurations. We test the following augmentation strategies:

- **Baseline**: SSA-Perin trained on the original non-augmented NoReC*fine* data.
- **Outside**: Using NorBERT2 to for generating new tokens, but only outside the spans of the original sentiment annotations. We also test various confidence thresholds, $p \in \{0.15, 0.5, 0.75\}$.
- **Inside**: Like above, but considering tokens that are inside the annotation spans only.

[2]https://github.com/ltgoslo/norsentlex

	p	DA-rate	Tuple	Targets	+/−
Baseline	n/a	0%	42.68	55.31	92.20
Outside	0.15	312%	41.60	54.71	91.89
	0.50	59%	**44.44**	57.23	92.07
	0.75	18%	43.72	56.88	92.32
Inside	0.15	190%	43.07	56.24	90.52
	0.50	33%	44.37	**57.24**	91.88
	0.75	9%	44.06	56.31	92.16
In/Out	0.75	27%	43.95	56.64	92.19
In/Out+Lex	0.75	21%	43.68	55.26	**92.66**
Upsampled	n/a	27%	43.33	56.77	92.02

Table 1: Results for various configurations on the NoReC$_{fine}$ *development* data. DA-rate corresponds to the percentage-wise increase in training sentences for a given MLM probability threshold p.

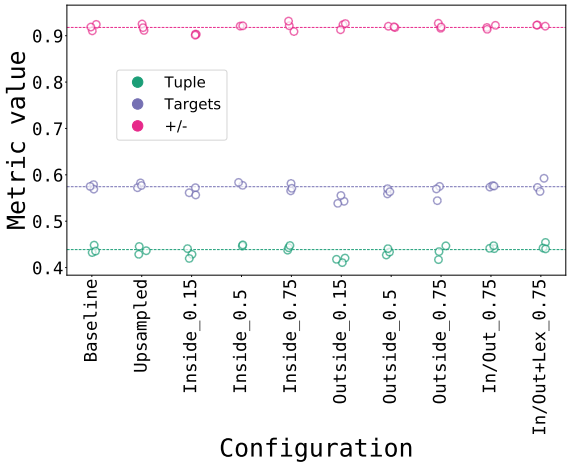

Figure 2: Showing the individual results for each of the 5 different runs for each configuration on the NoReC$_{fine}$ development data.

- **In/Out**: Generating new tokens for positions both inside and outside sentiment annotations. Note that we only do this for the threshold $p \geq 0.75$, as this has the lowest training time.
- **In/Out+Lex**: Like In/Out above, but additionally constraining augmentation by rejecting substitutions that change the *a priori* lexical positive/negative polarity for words listed in the NorSentLex sentiment lexicon.
- **Upsampled**: a control experiment where we duplicate the sentences in the training set with the same ratio as for the In/Out 0.75-threshold run, but without any token replacements.

Evaluation measures. We use three different evaluation metrics: (1) F1 score for the full sentiment-tuple, as described by (Samuel et al., 2022). This is a rather strict measure that takes into account the predictions of all expression spans – targets, holders, and polar expressions – in addition to the polarity. (2) Target F1 considers only the prediction of target expressions. Finally, (3) the +/− score is the accuracy of the positive/negative polarity prediction for the correctly predicted targets. For all configurations we run the models five times and report the averages.

Discussion of results. While we can see from the development scores in Table 1 that there are no huge differences in overall results, some trends are indeed noticable. We find that most of the augmented configurations perform better than the non-augmented baseline. However, our results so far do not allow us to conclude whether performing data augmentation only inside or outside of annotations, or both, works best. It is clear, though, that the confidence threshold of 0.15 is too low, especially when allowing substitutions outside of the sentiment annotations (which also yields the highest DA-rate), and the optimal threshold is likely somewhere in the higher range, between 0.5–0.75, and some further fine-tuning could be worthwhile.

It is interesting to observe the effect of adding the constraint that substitutions within the span of annotated polar expression are not allowed to switch the *a priori* word polarity with respect to a sentiment lexicon. Looking at the development results, this appears to indeed slightly boost the scores of the polarity predictions themselves. The same boost can not be seen for the held-out results in Table 2, however. Moreover, this constraint seems to reduce the scores for the target expressions and the full sentiment tuple, although this might potentially just be due to the corresponding reduced augmentation rate (given that certain substitutions are blocked). We believe an interesting direction for future work could be to implement more accurate strategies for detecting polarity-shifts during augmentation.

As can be seen in Figure2, there is some amount of variance for all configurations, making comparisons more difficult. Further tuning of the learning rate and regularization might be beneficial to reduce the variance. Note that we also tested a configuration where we merged all word substitutions for a given training example into a single sentence. The rationale for this was to test whether our stan-

	p	DA-rate	Tuple	Targets	+/−
Baseline	n/a	0%	43.39	54.13	92.59
Outside	0.50	59%	**45.08**	56.18	92.95
	0.75	18%	44.33	55.39	92.74
Inside	0.50	33%	43.19	55.76	92.46
	0.75	9%	43.38	55.62	92.49
In/Out	0.75	27%	44.12	**56.44**	**93.19**
In/Out+Lex	0.75	21%	43.66	55.53	92.55
Upsampled	n/a	27%	43.53	56.41	92.33

Table 2: Results for various configurations on the NoReC$_{fine}$ *held-out* test data. DA-rate corresponds to the percentage-wise increase in training sentences for a given MLM probability threshold p.

dard approach of multiplying out all substitutions in the augmented data, adding a near-duplicate instance of a training sentence for every word substitution, could cause overfitting. However, we found that for these runs the training would in several cases not converge (hence this configuration is not included in the table of results). Inspection also reveals that this approach more often end up generating semantically incoherent sentences. On the other hand, we see that the runs with the upsampled sentences yields quite robust results, with less variance than some of the DA-configurations, so overfitting from upsamplig effects does not seem to be an issue. Indeed, we see that the results of the upsampled runs very closely match those of the corresponding augmented runs, thereby indicating that some of the gains seen from MLM-DA may in fact simply stem from upsampling effects (e.g., by mitigating the possible undertraining of the baseline model).

Error analysis of augmented examples. Below we include some examples that show how seemingly minor changes to the original text can subtly impact polarity, potentially invalidating the original annotation. They also demonstrate some of the potential pitfalls of using MLM-based word substitutions for DA. When showing example sentences in the augmented data, we use the formatting **original/substitution** to indicate the original masked-out word and the generated substitution.

As a first case in point, the distributional word similarity captured by LMs will often lead a model to consider antonyms interchangeable, which in some contexts can reverse the polarity conveyed by the overall utterance, as in the example below:

(1) *En kollega av meg* ***sluttet/begynte*** *å se serien*

'A colleague of mine **stopped/started** watching the series'

In some cases, changing even just the tense of a verb can have subtle implications for the polarity, as in the example below:

(2) *Vi* ***har/hadde*** *har fortjent dette trofeet*

'We **have/had** deserved this trophy'

There also appears to be a slight tendency for closed-class words to be replaced, rather than content words. This is perhaps not so surprising, given that the probability for such replacements will likely be higher. However, these replacements also tend to surprisingly often nudge the polarity in new directions, as in the examples below:

(3) *Det er en interessant dokumentar ,* ***men/som*** *holder et svært høyt tempo.*
'It is an interesting documentary, **but/which** holds a very high pace.'

(4) *Ikke minst er det gøy å se det i 3D –* ***hvor/selvom*** *dansernes piruetter nesten pirker deg på nesa.*

'Not the least is it fun to watch it in 3D – **where/although** the dancers' pirouettes almost pricks you on the nose.'

Filtering words to be masked based on part-of-speech and/or frequency could perhaps be considered for countering some of the effects of the thresholding, shifting the augmentation more towards content words. We also find other examples that indicate that some tokens should perhaps be barred from masking and/or substitution, e.g. negation cues like *ikke* ('not').

6 Conclusion

We have presented a suite of experiments with using a pre-trained masked language model for augmenting the annotated training data for structured sentiment analysis. While we find that the described augmentation strategy often leads to improvements, the effects are modest, and high variance makes it difficult to draw definite conclusions. Our analysis of the results, based on a dataset of annotated Norwegian review data, point to several directions for further research on this topic, such as filtering candidate tokens for substitution with respect to frequency, PoS, polarity, negation cues, and more.

Acknowledgements

This work has been carried out as part of the SANT project (Sentiment Analysis for Norwegian Text), funded by the Research Council of Norway (grant number 270908).

References

Jeremy Barnes, Laura Oberlaender, Enrica Troiano, Andrey Kutuzov, Jan Buchmann, Rodrigo Agerri, Lilja Øvrelid, and Erik Velldal. 2022. SemEval 2022 task 10: Structured sentiment analysis. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1280–1295, Seattle, United States. Association for Computational Linguistics.

Jeremy Barnes, Samia Touileb, Lilja Øvrelid, and Erik Velldal. 2019. Lexicon information in neural sentiment analysis: a multi-task learning approach. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, Turku, Finland.

Markus Bayer, Marc-André Kaufhold, and Christian Reuter. 2022. A survey on data augmentation for text classification. *ACM Computing Surveys*, 55(7):1–39.

Cong Chen, Jiansong Chen, Cao Liu, Fan Yang, Guanglu Wan, and Jinxiong Xia. 2022. MT-speech at SemEval-2022 task 10: Incorporating data augmentation and auxiliary task with cross-lingual pretrained language model for structured sentiment analysis. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1329–1335, Seattle, United States. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for nlp. In *Findings of the Association for Computational Linguistics*. Association for Computational Linguistics.

Steven Y Feng, Aaron W Li, and Jesse Hoey. 2019. Keep calm and switch on! preserving sentiment and fluency in semantic text exchange.

Zhiting Hu, Bowen Tan, Russ R Salakhutdinov, Tom M Mitchell, and Eric P Xing. 2019. Learning data manipulation for augmentation and weighting. *Advances in Neural Information Processing Systems*, 32.

Sosuke Kobayashi. 2018. 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)*, pages 452–457, New Orleans, Louisiana. Association for Computational Linguistics.

Ashutosh Kumar, Kabir Ahuja, Raghuram Vadapalli, and Partha Talukdar. 2020. Syntax-guided controlled generation of paraphrases. *Transactions of the Association for Computational Linguistics*, 8:330–345.

Per Kummervold, Freddy Wetjen, and Javier de la Rosa. 2022. The Norwegian colossal corpus: A text corpus for training large Norwegian language models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3852–3860, Marseille, France. European Language Resources Association.

Andrey Kutuzov, Jeremy Barnes, Erik Velldal, Lilja Øvrelid, and Stephan Oepen. 2021. Large-scale contextualised language modelling for Norwegian. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 30–40, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of 7th International Conference on Learning Representations, ICLR*.

Lilja Øvrelid, Petter Mæhlum, Jeremy Barnes, and Erik Velldal. 2020. A fine-grained sentiment dataset for Norwegian. In *Proceedings of the 12th Edition of the Language Resources and Evaluation Conference*, Marseille, France.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv e-prints*.

David Samuel, Jeremy Barnes, Robin Kurtz, Stephan Oepen, Lilja Øvrelid, and Erik Velldal. 2022. Direct parsing to sentiment graphs. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 470–478, Dublin, Ireland. Association for Computational Linguistics.

David Samuel and Milan Straka. 2020. ÚFAL at MRP 2020: Permutation-invariant semantic parsing in PERIN. In *Proceedings of the CoNLL 2020 Shared Task: Cross-Framework Meaning Representation Parsing*, pages 53–64, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Erik Velldal, Lilja Øvrelid, Eivind Alexander Bergem, Cathrine Stadsnes, Samia Touileb, and Fredrik Jørgensen. 2018. NoReC: The Norwegian Review Corpus. In *Proceedings of the 11th edition of the Language Resources and Evaluation Conference*, Miyazaki, Japan.

Yiben Yang, Chaitanya Malaviya, Jared Fernandez, Swabha Swayamdipta, Ronan Le Bras, Ji-Ping Wang, Chandra Bhagavatula, Yejin Choi, and Doug Downey. 2020. Generative data augmentation for commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1008–1025, Online. Association for Computational Linguistics.