Targeted Sentiment Analysis Using a Fine-Grained Dataset: NoReCfine

Anonymous ACL submission

Abstract

Targeted sentiment analysis (TSA) is a variant of normal sentiment analysis, with which you can analyse the sentiment towards a specific target within a text. The goal of this project was to train different neural systems to perform TSA for Norwegian text. Different design methods were examined to find out what there effects were on the recently released data set $NoReC_{fine}$. The performance of bidirectional LSTMs, GRUs and a BERT model were compared to each other and a baseline. In addition a small error analysis was performed. The results show that BERT does not outperform GRUs or LSTMs.

1 Introduction

Sentiment analysis (SA) is a well-researched topic in Natural Language Processing (NLP). Conventionally the analysis obtains and classifies emotions or sentiments in sentences or documents as positive or negative. In this regard, there has been a parallel growth of SA identified with the growth of social media. This influx of information and with the use of a rich computational study as SA provides valuable information to marketers, managers or even manufactures on what consumers or employees perceive or demand (Xu et al., 2019). Thus, opinions derived from the study can change the policy of an organization as well as strategies to meet consumer demands. Nevertheless, SA is limited to deriving the overall polarity other than the reference aspect resulting in the development of SA branches such as targeted sentiment analysis (TSA) (Barnes and Klinger, 2019). Errors that may result from SA are harboured using TSA as the sentiment polarity identification (e.g., negative, neutral, or positive) is attached to a target in their context sentence (Pang and Lee, 2008; Jiang et al., 2011; Saeidi et al., 2016). For example, *"I hate his late coming behaviour, but he is an exceptional performer"* the reviewer has a negative sentiment toward an employee attendance behaviour but a positive sentiment on the performance. An aggregated sentiment drawn from the targets can allow the users to understand employee capability. This is achieved by the use of a neural network to automatically learn dimensional representation for targets in the sentence context. In this paper, we examine the effect of various design methods on $NoReC_{fine}$ dataset (Øvrelid et al., 2019) by identifying the target opinion and their associated polarity. With Bidirectional Long Short Term Memory (BiLSTM) as base model, we conduct experiments with other gated mechanisms like Gated Recurrent Unit (GRU) accompanied with deep neural layers.

The code and its documentation can be found on GitHub pages (Dujardin et al., 2020).

2 Related Work

Several machine learning methods have been used in TSA and despite its initial design with support vector machine (SVM), there has been an achieved progress by utilizing neural networks to encode base features as continuous and low dimensional vectors. Some studies employed Recursive Neural Network in conducting semantic composition for prediction representation while LSTM implemented in modelling the left and right context of a target and concatenate them for prediction results which were essentially not target specific (Jiang et al., 2011; Socher et al., 2011; Dong et al., 2014; Ma et al., 2018; Li and Lu, 2017; Tang et al., 2016).

The target influence on prediction representations were achieved by attention mechanism approach leading to bidirectional models that interactively learn attention weights on contexts and targets to separately generate target and context (Wang et al., 2016; Chen et al., 2017). In addition, more refined models are being studied to reduce

information loss in case of multiple words. For instance, to capture word-level interaction within a target and context, a mix of fine and course-grained attention mechanism was employed which could also depict the same context target interaction (Fan et al., 2018).

Transformer architecture modification introduces a lightweight model in TSA, which adopts attention-based encoders in modelling target words and context differently (Song et al., 2019). For sentences with more than one target word, (Zhou et al., 2019), utilizes Convolutional Neural Network (CNN) to model explicit dependency between the opinion words; the architecture obtains significant results on multiple target words in a sentence.

Regardless of the advantages of various designs, some studies show high state of the art results on certain dataset than others with regards to its composition (Chen et al., 2017) In addition, most of these architectures have been done on English datasets except a few with an example of a recent study that used fine-grained sentiment analysis on novel Norwegian dataset $NoReC_{fine}$ (Øvrelid et al., 2019). Our study will employ the same dataset to estimate the effect of computational TSA design as used on English dataset on the polarity and target of the data.

3 Data

3.1 NoReC dataset

The $NoReC_{fine}$ dataset is composed of a training, validation and testing subdatasets. The data contains labeled sentiment analysis of a variety of examples, from movies reviews to personal comments. The labels follow the Beginning - Inside - Outside (BIO) format, which are used to mark the start and end of a target within a sentence - positive or negative (Øvrelid et al., 2019).

Words	Labels
Men	O
den	B-targ-Negative
er	O
svakere	O
enn	O
”	B-targ-Positive
Sammen	I-targ-Positive
for	I-targ-Positive
livet	I-targ-Positive
”	I-targ-Positive
.	O

Table 1: Example of a sentence in $NoReC_{fine}$ dataset.

In addition, pre-trained word embeddings on a Norwegian-Bokmaal CoNLL17 corpus were used, obtained using a Word2Vec Continuous Skipgram model.

(a) Words inside negative targets.

(b) Words inside positive targets.

Figure 1: $NoReC_{fine}$ dataset. Words contained within positive or negative targets are relatively similar, however some patterns seems to differentiate these sentiments, like *love*, *unbroken* keywords.

3.2 Distribution

The dataset is composed of a total of 8.259 examples, splitted in training, validation and testing subdatasets.

	Train	Dev	Test	Total	Avg. Length
Sentences	6145	1184	930	8259	16.8
Targets	4458	832	709	5999	2

Table 2: Number of sentences and targets per datasets.

The labels are not equally distributed over the sentences. In fact, the dataset is imbalanced and 90% of the labels are the label *Outside* (O) of a target.

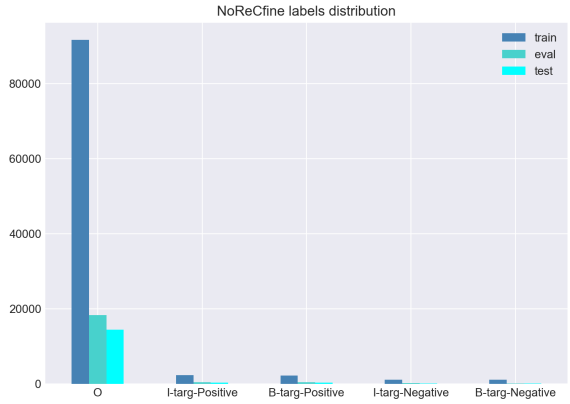


Figure 2: Labels distribution over the train, validation and test $NoReCfine$ datasets.

	Train	Dev	Test	Total
O	91691	18336	14425	124452
I-targ-Pos	2339	436	348	3123
B-targ-Pos	2245	433	366	3044
I-targ-Neg	1114	207	117	1438
B-targ-Neg	1094	196	145	1435

Table 3: Imbalanced labels distribution.

In addition, more than one sentiment can be retrieved from a sentence.

	Train	Dev	Test	Total
O	15.5	15.9	16.1	15.6
I-targ-Pos	3.3	3.0	3.0	3.2
B-targ-Pos	1.4	1.3	1.4	1.4
I-targ-Neg	3.1	3.0	2.3	3.0
B-targ-Neg	1.2	1.1	1.2	1.2

Table 4: Number of labels per sentence (mean).

4 Models

4.1 Approach

Our model is based on a provided baseline model (BiLSTM) for the TSA to analyse the polarity of the target words. As the script given did not conform precisely with (Paszke et al., 2019) standards, and the model was struggling to classify the task words, we further developed a compatible PyTorch version that had better results. The BiLSTM model was trained on Norwegian Bokmål word embeddings by utilizing random initialization with a one step forward pass and used cross-entropy loss function on an un-padded index to calculate the loss. In general LSTM models are capable of flexibly capturing the semantic relationship that exists between a target and the context words and avoids gradient vanishing (Tang et al., 2016).

To compare the results, a variant of an LSTM; Gated Recurrent Unit (GRU) model was developed. GRU introduces the combination of the input and forget gates resulting into a simplified gate. While among other changes the model also combines the hidden state to the cell state. As there exist mixed findings between LSTMs and GRU performance, the application of GRU on the $NoReC_{fine}$ data could establish an explainable result. The parameters remained virtually the same in both models.

Then, we wanted to explore the possibilities and results with a BERT model template. We preprocessed the $NoReC_{fine}$ datasets and generated masked arrays, but we faced the same problem of over-fitting.

4.2 Baseline

Our baseline model is a LSTM with a word embedding layer, where the weights were downloaded from NLPL vectors. We then used a hidden dimension of $H = 128$ for the LSTM cell, and made multiples variants. We explore different combinations of LSTMs, and then we moved on to a GRU base model.

We found that the GRU cell offers a better F1-score, but loss diverges after 40 to 50 epochs - an over-fitting issue.

4.3 BERT

Bidirectional Encoder Representations from Transformers (BERT) models are widely used for language modelling and sequence classification (Devlin et al., 2018).

For this short project, we used the $BERT_{BASE}$ multi-lingual pre-trained model on 104 language (latest model). We could have used the $BERT_{LARGE}$ which is made of a succession of 24 encoders (instead of 12), but we opted for a time-efficient model.

4.4 Loss Function

We first used a Cross Entropy Loss function, but a majority of models were over-fitting due to the imbalanced data. Thus, we tried a weighted loss which improved the results but did not solve over-fitting issues on baseline models.

The weights were determined from the distribution (table 3). We attributed a symbolic weight of 1 for the *PAD* label, but we ignored its effect when computing the loss items.

Target	Weight
O	0.0677
I-targ-Positive	0.9766
B-targ-Positive	0.9772
I-targ-Negative	0.9892
B-targ-Negative	0.9893
PAD	1.0000

Table 5: Weights used for the Cross Entropy Loss function.

5 Results

5.1 Over-fitting

The major issue we faced was over-fitting. Baseline models are not efficient when it comes to classify imbalanced data, and even though global accuracy converges, the loss (and F1-score) diverges.

5.2 Grid Search

To help us in hyper-parameters optimization, we created a custom package only for this purpose, inspired from *skorch*. We ran the hyper-search on a set of 200+ parameters combinations - weights, learning rate, LSTM/GRU cell's hidden dimension, dropout, batch size.

For BiLSTMs and BiGRUs, results are not really satisfying as loss diverge at epoch 20 maximum (Appendix, Figure 4). Even with weights, the model struggle to classify correctly under-represented targets (*B-targ-Negative* for example). However, as the class *O* is over-represented, global accuracy converge, no matter the parameters combination we used. The confusion matrices (Appendix, figure 5) highlight these issues - baseline models overfit the data even with a weighted loss function.

The grid search resulted in a variety of models, and we kept the best ones for each structure (Table 6). The BiGRU is our top F1-score model with the lowest loss.

5.3 Error analysis

Figure 5 shows that the most common errors for the BiGRU architecture are *Outside (O)* labels mismatched with actual meaningful target labels. The BiGRU seems to predict the label *O* for the majority of the time. The model seemed to The model also seems to struggle more with correctly predicting negative labels (*I-targ-negative* and *B-targ-negative*) compared to the prediction of positive labels (*I-targ-positive and B-targ-positive*). In addition, our BERT template didn't achieved more than a 0.12 F1-score with a respective 0.97 accuracy. As found by Xu et al. (2019), BERT weights may work well with some NLP tasks but fail with review-based tasks

We didn't have enough time to change or try different BERT structures, like RoBERTa with different folds that seems to work well for sentiment extraction - our main issue.

Last but not least, the dataset is more complex to handle as more than one sentiment can be extracted from a sentence. As there is only one *B-targ-Sentiment* label for a sentiment, if the begining is mismatched the F1-score will drop significantly. We tried to add different weights to tackle this issue, with a significant improve (+0.1 F1-score approximately), but our models remained under the 0.5 score threshold.

The imbalanced dataset causes the model to over-

Target	Loss	Accuracy	F1 score	Recall
BiLSTM	0.4582	0.8889	0.2884	0.3527
BiGRU	**0.2415**	**0.9371**	**0.3501**	**0.3451**
BERT-base-multilingual-cased	0.3106	0.9717	0.1157	0.1657
BERT-large-multilingual-cased	0.3320	0.9602	0.0354	0.0152

Table 6: Results of best models from different structures, evaluated on the validation dataset.

fit the data, and only the class *O* (outside) is correctly predicted with a satisfying accuracy on the validation dataset.

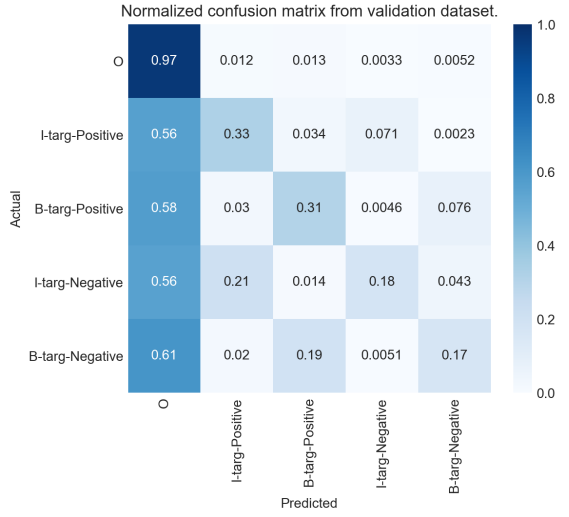


Figure 3: Confusion matrix obtained using our BiGRU recurrent network, trained over 100 epochs.

5.4 Conclusion

Targeted Sentiment Analysis is a variant of Sentiment Analysis. In TSA the sentiment towards a specific target in a text is analysed. Starting with a simple baseline model, a gridsearch was performed with a BiLSTM and BiGRU to find out optimal parameters. The results showed that the models are suffering with overfitting. The global accuracy remained high (around 0.9), the loss diverged. A possible explanation for this might be that the model predicts the majority class ('label *O*') and therefore correctly classifies a majority, but also missclassifies a lot of data therefore increasing the loss.

Targeted Sentiment Analysis is a similar topic to Aspect Based Sentiment Analysis, but with a more complex sentiment extraction. Even if opinion mining is a solved research subject, the uncertainty of a starting and ending target makes TSA predictions much more difficult to solve. We figured that simple models (RNN-like) perform better than more complex ones (BERT).

BERT models are well used in Named Entity Recognition areas, and we were convinced that it could outperforms our baseline model. However, because starting tokens depend mainly on the context (and not necessarily on the words) BERT models predict more *O* than reality, leading to a decrease in F1-score.

We did not have enough time to try different tuning of BERT, like a RoBERTa or DistilBERT, which are recently used and may perform better with $NoReC_{fine}$ dataset. For the future we still believe a BERT model might elicit a good performance. Implementing a BERT model is something that could be explored in further research.

References

Barnes, J. and Klinger, R. (2019). Embedding projection for targeted cross-lingual sentiment: Model comparisons and a real-world study. *CoRR*, abs/1906.10519.

Chen, P., Sun, Z., Bing, L., and Yang, W. (2017). Recurrent attention network on memory for aspect sentiment analysis. pages 452–461.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding.

Dong, L., Wei, F., Tan, C., Tang, D., Zhou, M., and Xu, K. (2014). Adaptive recursive neural network for target-dependent twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 49–54, Baltimore, Maryland. Association for Computational Linguistics.

Dujardin, A., Boerboom, L., and Hertzberg, S. (2020). Sentarget: A python package for targeted sentiment analysis.

Fan, F., Feng, Y., and Zhao, D. (2018). Multi-grained attention network for aspect-level sentiment classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3433–3442, Brussels, Belgium. Association for Computational Linguistics.

Jiang, L., Yu, M., Zhou, M., Liu, X., and Zhao, T. (2011). Target-dependent twitter sentiment classification. pages 151–160.

Li, H. and Lu, W. (2017). Learning latent sentiment scopes for entity-level sentiment analysis. In *AAAI*.

Ma, D., Li, S., and Wang, H. (2018). Joint learning for targeted sentiment analysis. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2:1–135.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019).

Pytorch: An imperative style, high-performance deep learning library. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Saeidi, M., Bouchard, G., Liakata, M., and Riedel, S. (2016). Sentihood: Targeted aspect based sentiment analysis dataset for urban neighbourhoods.

Socher, R., Lin, C., Ng, A., and Manning, C. (2011). Parsing natural scenes and natural language with recursive neural networks. pages 129–136.

Song, Y., Wang, J., Jiang, T., Liu, Z., and Rao, Y. (2019). Targeted sentiment classification with attentional encoder network. *Lecture Notes in Computer Science*, page 93–103.

Tang, D., Qin, B., Feng, X., and Liu, T. (2016). Effective LSTMs for target-dependent sentiment classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3298–3307, Osaka, Japan. The COLING 2016 Organizing Committee.

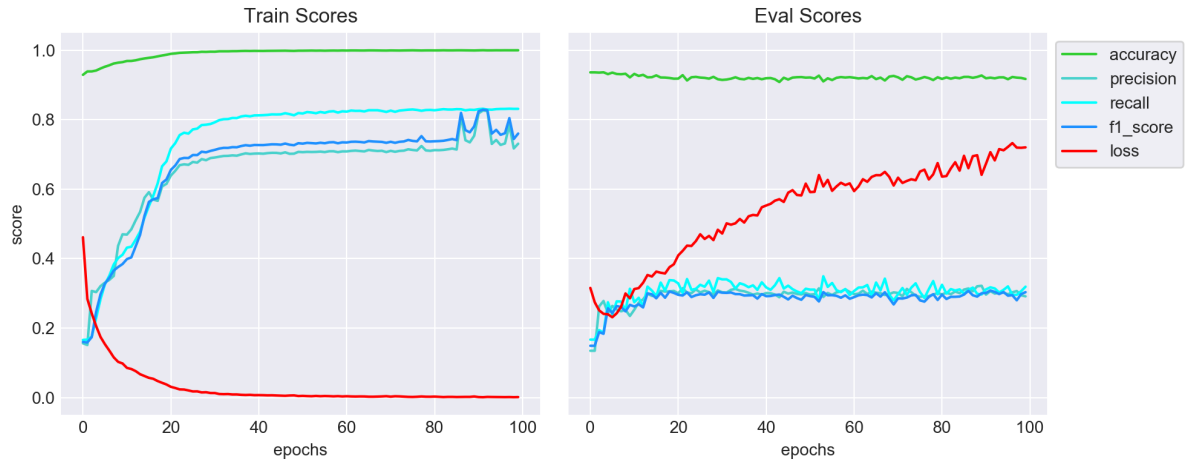
Wang, J., Shen, X., Sun, Y., and Qu, A. (2016). Classification with unstructured predictors and an application to sentiment analysis. *Journal of the American Statistical Association*, 111(515):1242–1253.

Xu, H., Liu, B., Shu, L., and Yu, P. (2019). *BERT Post-Training for Review Reading Comprehension and Aspect-based Sentiment Analysis*.

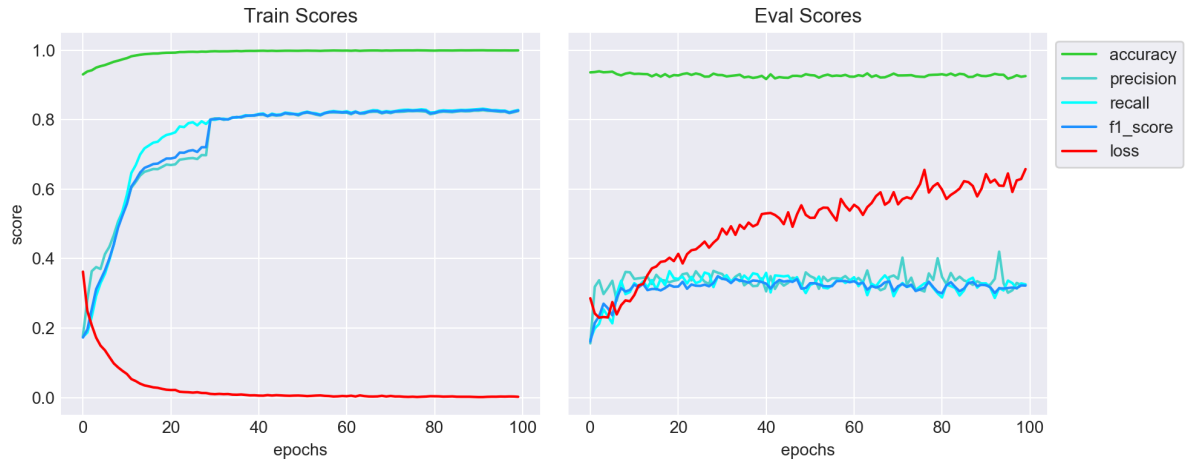
Zhou, X., Zhang, Y., Cui, L., and Huang, D. (2019). Evaluating commonsense in pre-trained language models.

Øvrelid, L., Mæhlum, P., Barnes, J., and Velldal, E. (2019). A fine-grained sentiment dataset for norwegian.

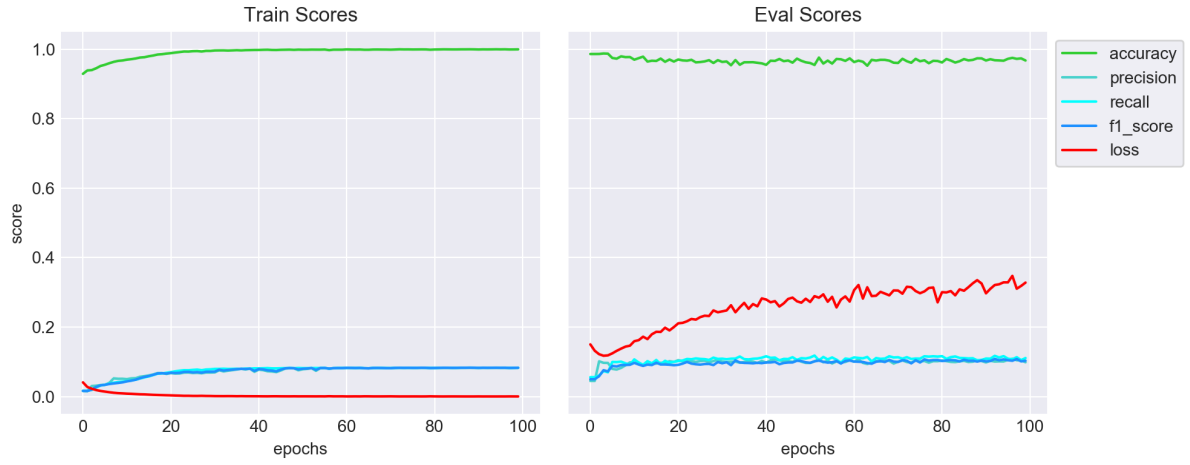
Appendices



(a) BiLSTM results.

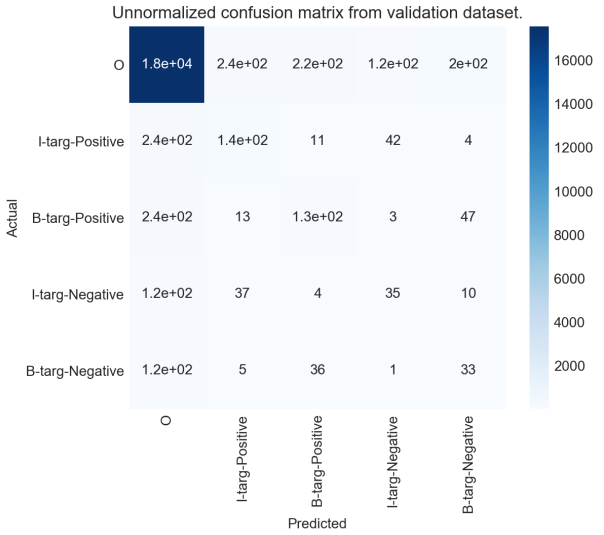


(b) BiGRU results.

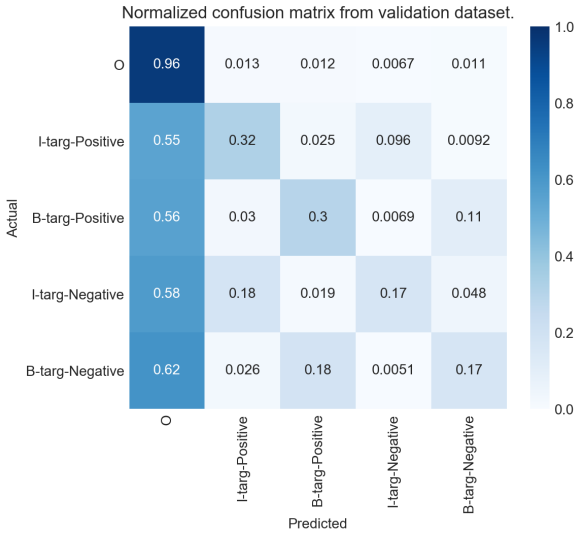


(c) BERT-base-multilingual results.

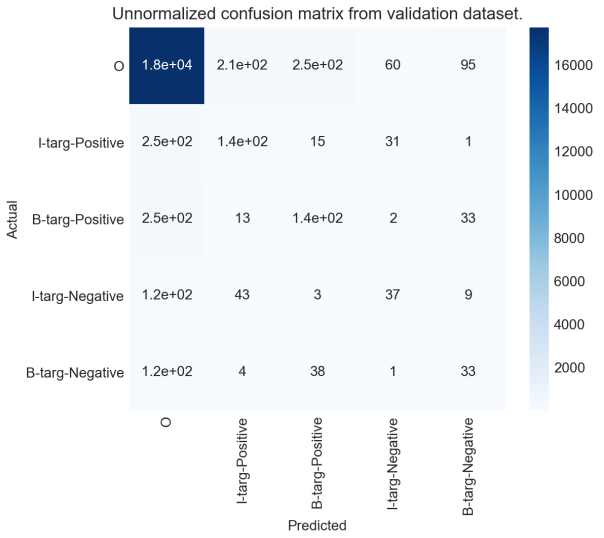
Figure 4: Results of best models from different structures, evaluated on the validation dataset. Even though the accuracy converge in all cases, loss increases when models are training for too long. The shift occurs relatively soon (5-20 epochs), and we experienced that it happens regardless of our hyper-parameters tuning.



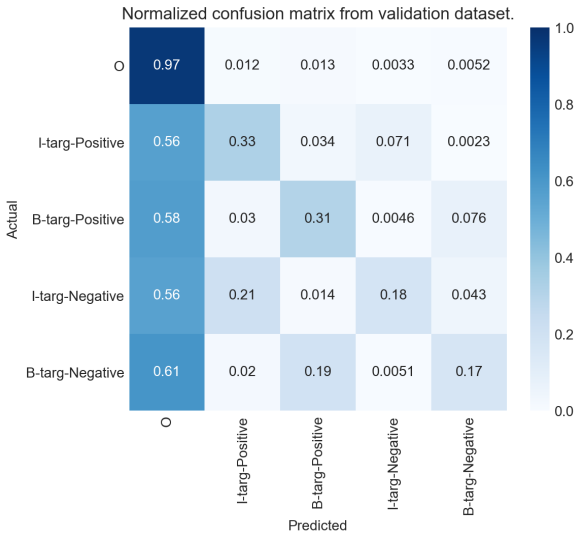
(a) Unnormalized BiLSTM confusion matrix.



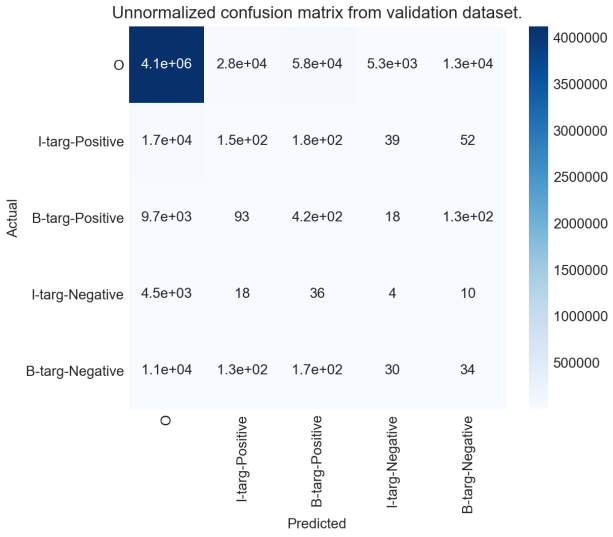
(b) Normalized BiLSTM confusion matrix.



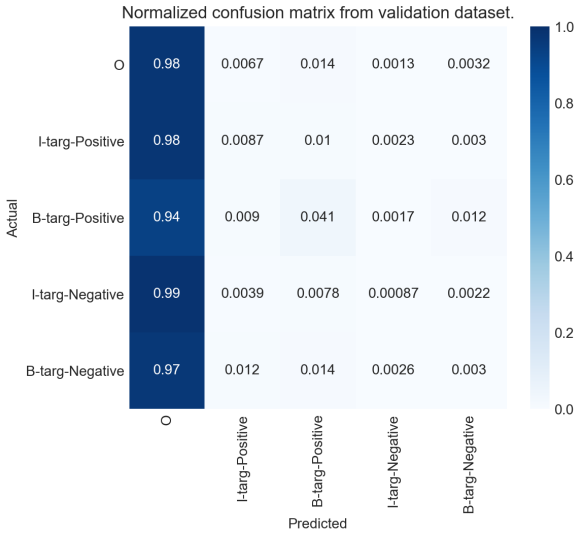
(c) Unnormalized BiGRU confusion matrix.



(d) Normalized BiGRU confusion matrix.



(e) Unnormalized BERT confusion matrix.



(f) Normalized BERT confusion matrix.

Figure 5: Unnormalized (left) and normalized (right) confusion matrices of best model taken from each structures, evaluated on the validation dataset.