Metaphor Detection using the Kontextbruch approach in a Multi-Task Learning Setup

Dennis Holzmann
dennis.holzmann@uni-bielefeld.de
Fynn Martin Gilbert / 3093453
fynn.gilbert@uni-bielefeld.de

Abstract

Detecting metaphors has been a challenging task till this day. In this paper we trained two BERT models in order to detect metaphors in old German texts. For that, a dataset is used which has been annotated according to the Kontextbruch concept. But instead of using just the gold label standard, we made use of the soft-labels by introducing a multi-task learning setup where the main task is to classify a sentence as metaphor or non-metaphor and the secondary task is to predict the soft label of that sentence which is the probability distribution over the labels given by the annotators. Additionally, we added the task of estimating the emotions behind the sentence as a third task. For this, we use a BERT model pretrained on old german texts and a with the VUA data set intermediate trained MBERT model.

1 Introduction

Metaphors allow us to enhance the meanings of what we want to say. With them, it is possible to convey abstract topics and emotions in a way that is difficult or even impossible to describe without them. We use them everyday often without realising. They have become an essential part of our language. Nonetheless, detecting metaphors is still a very challenging task in the NLP community.

This paper aims to continue the works of (Nguey et al., 2021), where they evaluated the usage of BERT and MBERT and the effect of pretraining the model on several for metaphor detection annotated data sets before training on the Kontextbruch data set. We are trying to improve the results of metaphor detection by introducing multiple data sets into the training phase in a Multi-Task-Learning setup, that are not necessarily annotated for metaphor detection. In order to detect metaphors, we are using a corpus of old German texts which has been annotated using an approach called Kontextbruch (Gehring, 2010) which means ”context breach”. The main idea is, that a metaphor is breaching the context in a sentence in order to emit certain emotions or to convey abstract information. The context of a sentence is described as the **frame** while the metaphor that ”breaches” the frame is described as the <u>focus</u> (See excerpt 1). This concept differs from other metaphorical concepts as the conceptional metaphor theory where a metaphor is seen as a mapping of a source domain to a target domain (Lakoff, 1980).

Der **politische Körper** verwendet in beiden Fällen seine meiste Kraft auf die Zunahme von <u>Zähnen und Krallen</u>.

Table 1: Excerpt from the Kontextbruch data set. The <u>focus</u> is underlined while the **frame** is in bold.

Most approaches to metaphor detection only make use of the gold standard. Not always do annotators agree. The gold standard is the majority vote of the annotators for a sentence, while soft-labels are described to be probability distribution over the labels. We took inspiration in (Fornaciari et al., 2021), where they used the disagreement of the annotators in order to reduce overfitting. For that they set up MTL models where the auxilliary task would be to predict the label distribution. They found, that predicting the soft labels reduce overfitting.

Furthermore, metaphors often do contain strong emotions, as metaphors are often used to enhance emotions. Non-literal language induces a higher emotional response than using literal language (Fainsilber and Ortony, 1987; Mohammad et al., 2016; Crawford, 2009). We took inspiration from (Dankers et al., 2019), and added the task of emotion regression with the EmoBank data set (Buechel and Hahn, 2017).

The approaches described in (Dankers et al., 2019;

Fornaciari et al., 2021) showed improving results by introducing the MTL-setup. This paper focuses on employing these approaches to improve the detection of metaphors.

2 Related work

BERT The model that we use is the BERT and MBERT model (Devlin et al., 2019). The BERT model (Bidirectional Encoder Representations from Transformers) is a NLP model. As the name already suggests, BERT is based on Transformers and Multi-Head Attention (Vaswani et al., 2017). They way BERT is trained is essential for the success in NLP tasks. By employing two training methods the BERT model is pre-trained to learn to understand language and context. The first method is the Masked Language Model (MLM). A word in the input sentence is masked and the BERT model has to predict the missing word. The second method is the Next Sentence Prediction (NSP). After receiving one sentence as input, BERT has to predict what the next sentence out of two would be the sentence succeeding the input sentence. The BERT model output token representations which then can be used as input for other classification or regression models. Many pre-trained BERT model already exist and can be used to either directly fine-tune them or further pre-train the models[1]. We have been using a BERT model pretrained on a large corpus of old German texts, as the Kontextbruch data set contains German sentences. Based on the results of (Nguey et al., 2021), we chose to additionally use the MBERT model that has been pre-trained on the VUA data set (Steen et al., 2010). The MBERT model is a multilingual BERT model which has been trained on data in different languages, allowing us to use data sets for training that do not use the same language.

Multi-Task Learning The idea of Multi-Task Learning (MTL) is that learning multiple related tasks simultaneously and with shared parameters, the model should be able to better generalize on the original task (Caruana, 2004). Two main methods in parameter sharing are being used. With **hard-parameter-sharing** each tasks has its own task-specific layers while sharing hidden-layers. The hidden-layers in our case belong to the BERT model. When employing soft-parameter-sharing, each task has its own model, connected with the other task-specific models.
Each task can have different loss functions and can be trained on different data sets. A common way of combine the loss functions of each task is by taking the weighted sum of each loss. This way, it can be avoided that the auxiliary tasks take over and worsen the results by leading the gradient in an unfavourable direction. Because of the different data sets, the training for each task has to be scheduled in some way. Optimizing the task scheduling can highly improve the performance of the model (Bengio et al., 2009). In this paper, we choose to randomly select one of the task for each training step.

3 Data

In this chapter we introduce the dataset and the challenges that come with it. Since we have a low quantity of samples, we decide to ensure data quality through several preprocessing steps. This later allows us to increase data quantity by repeating the oversampling by translation technique that had been used by the previous group. This builds the data foundation for the following chapters.

The dataset we work with was provided by a previous student group from TU Darmstadt via github. Since the task is to find metaphors in old German texts, the data sets consists of Textstellen entries with one of three metaphor labels, ”Metapher” (3.1%), ”Metaphernkandidat” (13.7%) and ”Nein” (83.2%). As can be seen in figure 1, the classes are highly imbalanced.

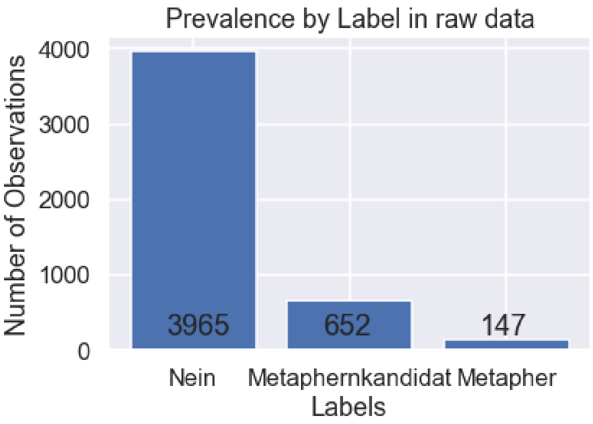


Figure 1: Class Prevalence

Additional information about the annotators and the potential ”frame” and ”focus” are also included among other features. There are three data sets in total. The first data set consists of only non-

[1] https://huggingface.co

metaphor entries. The other two data sets contain mostly the same Textstellen with metaphors or metaphor candidates. The difference between the latter two data sets is that one contains gold standard labels, while the other contains different labels for repeating Textstellen by differing annotators. It contains all gold standard entries and an additional 142 Textstellen. This allows us to first group the latter data set by Textstelle and aggregate the annotations. The grouped data can then be merged with the gold standard to achieve both soft and hard labels for each Textstelle. Since there are more than 799 aggregated Textstellen and only 657 gold standard ones, the missing 142 hard labels can be imputed via the silver-standard labels, aka the most frequent annotation. After the data pre-processing and augmentation steps, the tables are merged into a single data frame which forms the basis for training and evaluation.

3.1 Preprocessing

The provided data sets were built from several documents labeled by different annotators. This information was stored in different files with differing file types. The files were then read and aggregated via script into a single .csv-file. This had the effect that some artefacts made it into the final version of the table. Since we have only a small quantity of data for a big data task, we decide to focus on data quality. In order to allow for a high enough signal to noise ratio, we clean the Textstellen data in several steps:

3.1.1 Dropping duplicate entries

In order to check for duplicate texts in the non-metaphors and metaphors, sentence embeddings are created and compared via cosine similarity, the highest of which are candidate pairs for repeating entries. To create the sentence embeddings, a multilingual SBERT variant, distiluse-base-multilingual-cased-v2, is used (Reimers and Gurevych, 2020). Based on the embeddings, a cosine-similarity matrix is created. Going through the lower triangle of matrix entries along with the corresponding Textstellen allows for a quick identification of duplicate entries, that differ only in minor artefacts. Plotting the entries as in figure 2 allows for a visual identification of a reasonable threshold, above which the text pairs are passed on to manual inspection for false positives. Duplicates are usually removed by dropping the shorter entry, and in case of metaphor candidates and metaphors, adding the soft labels to the longer, remaining entry. For non-metaphor data, 60 duplicates are removed, while 7 metaphors or candidates are dropped.

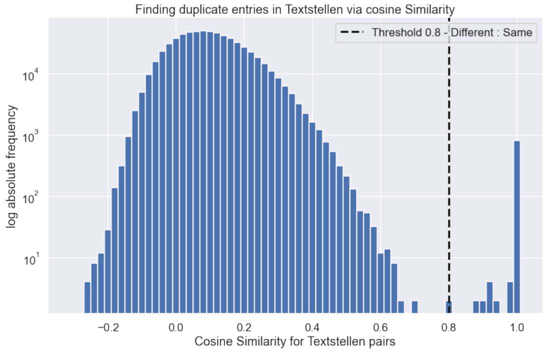


Figure 2: Cosine Similarities for Metaphor (Candidate) Textstellen

3.1.2 Removing unnecessary characters or fragments

While not indispensable, it is common practice to sanitize model inputs to train on a more uniform data set free of unnecessary noise. Characters like German quotation markers or line-breaks are removed, that remain from the original print. Other Textstellen begin with numbers or roman numerals, since they stem from enumerated passages, which are also cleaned. While these steps might not actually affect model performance there is one error that might affect results, if not addressed. In more than 20% of texts, there are repeating text fragments within a sentence, sometimes up to four times. The following Textstelle shows an extreme example:

"Übrigens haben wir bereits im 9. Kapitel (S. 235-240) gesehen, daß einerseits die Rechenmanöver, die sich auf das Glück einer Gesellschaft beziehen, seits die Rechenmanöver, die sich auf das Glück einer Gesellschaft beziehen, mit ebenso unüberwindlichen Schwierigkeiten zu kämpfen haben wie der Versuch, einen Regenbogen zu besteigen, und daß andererseits von dem Aus such, einen Regenbogen zu besteigen, und daß andererseits von dem Aus -sichtspunkt, den die Descendenztheorie gewährt, wahrnehmbar ist, daß Glücks sichtspunkt, den die Descendenztheorie gewährt, wahrnehmbar ist, daß Glücks -empfindungen irgendwelcher Gruppe oder Summe von Personen nicht das empfindungen irgendwelcher Gruppe oder Summe von Personen nicht das Ziel der inneren Politik sein können, sondern ausschließlich die Anpassung an die Bedingungen, von denen für das Gemeinwesen der Sieg im Daseinskampf abhängt."

Since these don't occur in the non-metaphor data, it is crucial to remove the fragments. Otherwise the model might not learn to identify metaphors and candidates, but instead only learns to identify these repeating fragments and mistakes them for metaphors.

4 Methods

4.1 Data augmentation - Oversampling by translation

While the preprocessing so far focuses on data quality, we now address data quantity. The technique that has been used by the previous group is oversampling by translation. It has shown promising results because it addresses the low number of training instances. Only translating rare classes (Candidates and Metaphors) mitigates the imbalanced nature of the original labels (see figure 3 or table 2).

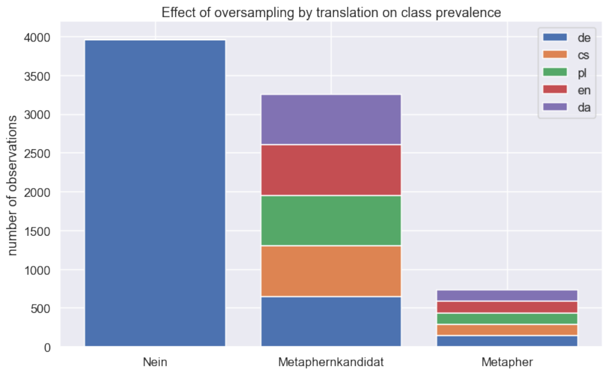


Figure 3: Class Prevalence after Backtranslation

	before	after
Metapher	147 (3.1%)	735 (9.2%)
Kandidat	652 (13.7%)	3260 (41.0%)
Nein	3965 (83.2%)	3965 (49.8%)

Table 2: absolute and (relative %) class frequency before and after oversampling by translation

By translating a Textstelle into another language and back into German, the hope is that exact words will slightly change, while the meaning of the sentence and the use of metaphors will stay largely the same, like in a game of Chinese whisper. This way we propagate the soft and hard labels to the translated sentence and create a new data point, on which we can train the model. The previous cleaning steps also help with the quality of the translation because the previous errors could propagate and might increase in the oversampling step.

Our expectation was that the sentence translations only change to a small degree in embedding space, and that this would only increase the weight of the original sentence in the training. When visualizing the embeddings as in figure 7 however, one can see that the translations tend to increase the sentences embedding neighborhood and can almost overlap with other sentences, at least when represented in two dimensional space. This promises to help the model generalize better.

To test if metaphors actually remain, we use the Dansk translations to get an estimate. We classify them manually into three categories and give examples:

- **Metaphor or meaning lost:**

-
  - original:
    "Ein ganz ähnliches Schicksal war alsdann den Römern beschieden, die Zügel der Weltherrschaft entsanken ihnen, jüngere Völker kamen obenauf."
  - after back translation:
    "Ein ganz ähnliches Schicksal ereilte die Römer; die Weltherrschaft entglitt ihnen, jüngere Menschen stiegen an die Spitze auf."

- Note the loss of the metaphorical use of the word "Zügel". In addition "Völker" (peoples) has been changed to "Menschen" (humans), which changes the meaning of the sentence.

- **Wording changed:**

- In other cases the wording changes, but the use of metaphor stays the same. This should help the model to generalize better. Since translations are modern versions of a language, the generalization might be across time, not across metaphors per se.

-
  - original:
    "Unerkannt sitzt die geschlechtliche Zuchtwahl aber auch am Webstuhl der Geschichte."
  - after back translation:
    "Unbekannte, sexuelle Selektion sitzt auch im Netz der Geschichte."

Concrete numbers can be seen in figure 4, where metaphors are lost in slightly less than 25 %.

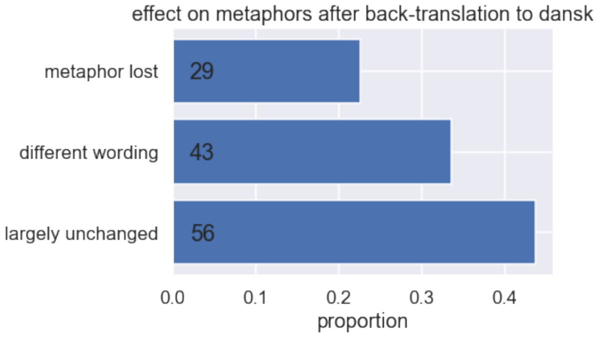


Figure 4: Effect of dansk back translation on gold standard metaphors

Oversampling by translation helps with the issues that come with a small sample size and class imbalance. As seen in the Dansk translations, it is not guaranteed that features remain, so that manual inspection is necessary.

We translate all metaphor or candidate Textstellen using google translate via the deep translator API. We decide, somewhat arbitrarily, to translate the sentences into english (en) and languages spoken in countries that border on germany: dansk (da), polish (pl) and czech (cs). Note that when using a Multilingual BERT variant, not just the sentences translated back into german can be used for training, but the original translation in different languages, if supported, can be used as well.

4.2 BERT and Variants

BERT In this paper, we used two models. The first model is a on german corpus pretrained BERT model which has been named ”redewiedergabe”. As the Kontextbruch data set contains German sentences, this model seemed to fit perfectly.

MBERT The second model that we have trained is MBERT which has been intermediate-task fine-tuned with the VUA training data set. The VUA data set contains English sentences that have been annotated for token-wise metaphor detection. With intermediate-task fine-tuning we hope to be able to use the large corpus to efficiently and more easily train the model with the much smaller Kontextbruch data set, which presumably on its own won’t be sufficient enough. The VUA does not contain the labels of each annotator respectively but only the gold standard.

MTL setup Using the soft-labels extracted from the Kontextbruch data set, we trained the redewiedergabe model with sentence classification (I.e. metaphor detection) as the main task and soft-label prediction as the auxiliary task.
We did the same with the fine-tuned MBERT model but additionally added a third task which tries to regress the emotions in the sentence using the EmoBank data set (Buechel and Hahn, 2017). The emotion in a sentence have been annotated according to the Valence-Arousal-Dominance (VAD) representation. The values for valence, arousal, and dominance range from 0 to 1. This data set contains English sentences and therefore we can’t use this dataset to train the redewiedergabe model.
We employ hard-parameter sharing for the MTL setup. Each task shares one BERT model and has its own task-specific layers. The authors of (Fornaciari et al., 2021) found that the inverse Kullback-Leibler Divergence loss results in the best performance. We therefore adopt the loss function for predicting soft-labels. For the emotion regression task, we adopted the RMSE-loss that has been used in (Dankers et al., 2019). Our MTL-Task setup divers from (Dankers et al., 2019) in the way that the lasts layers in our the BERT model are shared by the tasks while that is not the case in (Dankers et al., 2019).

4.3 Performance Metric

So far the performance metric that has been used is the F1-score. It is chosen because in cases of (severe) class imbalance, metrics like accuracy weight all classes equally. The F1-score and its constituents, recall and precsision, focus on the rare cases only, while metrics like specificity completely ignore the rare positive cases.

In our experiments we encounter the problem that our trained models don’t predict a single metaphor at all, and all models receive a F1-score of zero for the metaphor class.

However, a model that misses the correct metaphor prediction only by a single percentage point is obviously better than a model that misses the metaphor completely. Instead of manually lowering the metaphor threshold arbitrarily to find at least some metaphors, we decide to use a threshold independent metric.

Usually the ROC-AUC score is used in such scenarios, because it is a model wide metric. We use the average-precision-score (AVG-PRC) however, because it addresses the class imbalance problem by switching specificity for precision. It also is threshold independent and therefore combines the desirable traits of the F1 score and the ROC-AUC-score.

Even if two different models predict the same labels, it still shows a difference in quality on a probability prediction level. Since our data set is small, no computational performance issues arise.

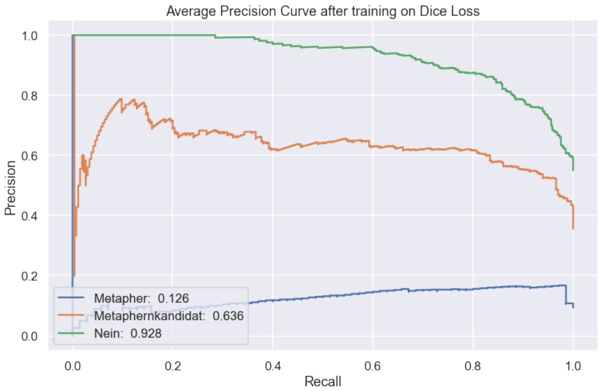


Figure 5: Average-Precision curves and scores for Redewiedergabe BERT model trained on Dice Loss

4.4 choice of loss function

While cross entropy is the de-facto standard loss function for classification, it can be insufficient when dealing with highly imbalanced datasets, for similar reasons as mentioned in section 4.3. There exist loss functions which take the class imbalance into account, like Dice Loss. In addition to vanilla cross entropy, we also use the focal loss (FL) function for training. FL has been introduced by Lin et al. (2017), to combat class imbalance in dense object detection. It can be seen as an extension of regular cross entropy, which down weights easy cases in order to focus on the hard cases:

$$FL(p_t) = -\alpha_t \sum_{k=1}^{K} (1 - p_{tk})^{\gamma} \log(p_{tk})$$

$$K: \text{ number of classes}$$

The modulating factor $(1 - p_{tk})^{\gamma}$ has the effect that hard cases are weighted higher, while the focusing hyperparameter $\gamma \geq 0$ determines how strongly the weights apply. For $\gamma = 0$ the formula defaults to balanced cross entropy. Since metaphors are generally very rare, not just in our dataset, we hope that FL helps to combat the class imbalance problem.

5 Experiments + Results

This section showcases our final model, including its performance and the necessary steps that lead to it. First we explain the trails that we use to replicate previous performance results. We reduce their success to it's main component via process of elimination. We use the insights to train our own BERT Variant and summarize its performance.

5.1 Repeating experiments from previous group

After cleaning and oversampling our dataset, we are unable to reproduce the results of the previous group. We go through several steps to pinpoint how they got there and how we can build on their work.

Repeating text fragments only occur in the metaphor and candidate examples. Therefore we hypothesise that the cleaning steps remove any metaphor information found before. That would mean, that in previous work, the group did not actually find metaphors, but instead found repeating text fragments, and mistook them for metaphors.

To test this hypothesis, we train the Redewiedergabe Model on both clean and unclean training data and for each model check their respective performance on both clean and unclean train and test data.

<table>
<tr><th>Evaluated on</th><th>unclean model</th><th>clean model</th></tr>
<tr><td>unclean train</td><td>16.7 | 53.5 | 90.7</td><td>05.3 | 52.9 | 92.0</td></tr>
<tr><td>clean train</td><td>17.7 | 65.3 | 77.9</td><td>08.4 | 68.1 | 87.8</td></tr>
<tr><td>unclean test</td><td>08.3 | 56.0 | 89.7</td><td>03.3 | 53.1 | 92.9</td></tr>
<tr><td>clean test</td><td>0.93 | 71.2 | 78.7</td><td>04.8 | 75.3 | 88.7</td></tr>
</table>

Table 3: Average Precision Scores for | Metaphors | Candidates | Non-Metaphors |, respectively. The Model was either trained on clean or unclean data and each is evaluated on training and test sets

As you can see in table 3, the model trained on cleaned data generalized better for the non metaphors and candidates, while unclean performs better on metaphors. However these results don't confirm our hypothesis.

The next hypothesis is that the high scores achieved prior are because original sentences and back-translations were mixed between training and testing set. This would mean that the model remembers labels by memorizing similar sentence embeddings instead of learning what metaphors are. However, the previous group paid attention to this detail by translating only training sentences on the fly. Therefore, this hypothesis also can't account for the difference in performance between the results.

This leaves us with the hypothesis that we could not repeat the results because of the down sampling technique used be the previous group. Down sampling comes with the problem that it reduces the already small data set. While it aims to reduce or eliminate class imbalance, the actual problem is only masked from the model. The hard part is to learn what the rare cases look like, despite the noise generated by the prevalent class. Therefore we did not aim to reproduce the results achieved by down sampling. However those results show already that there is information in the Textstellen that is picked up by BERT variants, which can be drowned in noise.

Instead, we try to tackle the class imbalance problem by other means. We use focal loss as minority class sensitive loss function. Experiments as in figure 6 show that this addresses the problem of class imbalance.

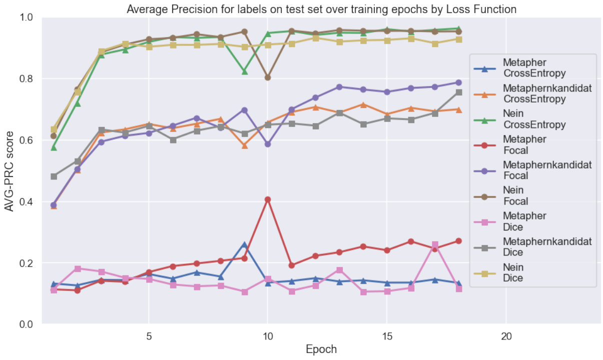


Figure 6: Average Precision scores for different Labels and Loss Functions (note the red and purple lines) on test data (10%)

We have shown that the main challenge for the model is class imbalance. With this finding we can extend and improve on the previous results.

5.2 MTL Experiments

We have trained the metaphor classification with the focal loss. Instead of having a ternary classification problem, we chose to treat each metaphor candidate as non-metaphor, creating a binary classification problem, as we suspect that the metaphor candidates would rather confuse the model. The metaphor detection head consists of a drop-out layer with a probability of 0.1, and linear layer with sigmoid activation for classification. The auxiliary task heads are strucutred similarly.

Accuracy	0.907
Precision non-Metaphor	0.907
Precision Metaphor	0.0

Table 4: Accuracy and Precision of the redewiedergabe model trained with soft-label prediciton as auxiliary task. The weights for each task is set to be equal.

Accuracy	0.897
Precision non-Metaphor	0.897
Precision Metaphor	0.0

Table 5: Accuracy and Precision of the redewiedergabe model trained with soft-label prediciton as auxiliary task. The soft-label task has been multiplied with 0.1.

5.2.1 Redewiedergabe

We have trained the redewiedergabe BERT model, which has been pre-trained on a large corpus of German texts. The first model has been trained without weighting the loss functions (see table 4). Even though, we expected the auxiliary task to avoid overfitting, the model rather learns to predict most of the sentences as non-metaphors. increasing the gamma value $\gamma = 5$ in the focal loss function did not result in an improvement. At first we expected the auxiliary task to be having too much control over the gradient, as the loss for generated by the auxiliary task was 10-100 times higher than the loss generated by the main task. We therefore weighted the auxiliary task by multiplying the loss with $w_{auxLoss} = 0.1$ (see table 5). Unfortunately, this did not lead to better results. Our BERT model has not been able to predict a single metaphor.

5.2.2 MBERT

The MTL setup with the MBERT model is similar to the setup used for the redewiedergabe model. However, with the MBERT model we were able to utilize the EMOBank data set in order to add an additional auxiliary task, which is trained to regress emotions from the sentence. The first two models have been trained without weighting the loss functions.

Accuracy	0.899
Precision non-Metaphor	0.899
Precision Metaphor	0.0

Table 6: Accuracy and Precision of the MBERT model trained with soft-label prediciton as auxiliary task. The weights for each task is set to be equal.

Accuracy	0.901
Precision non-Metaphor	0.901
Precision Metaphor	0.0

Table 7: Accuracy and Precision of the MBERT model trained with soft-label prediciton and emotion regression as auxiliary tasks. The weights for each task is set to be equal.

Accuracy	0.899
Precision non-Metaphor	0.899
Precision Metaphor	0.0

Table 8: Accuracy and Precision of the MBERT model trained with soft-label prediciton as auxiliary task. The auxiliary tasks have been multiplied with $w_{auxLoss} = 0.1$.

Accuracy	0.901
Precision non-Metaphor	0.901
Precision Metaphor	0.0

Table 9: Accuracy and Precision of the MBERT model trained with soft-label prediciton and emotion regression as auxiliary tasks. The auxiliary tasks have been multiplied with $w_{auxLoss} = 0.1$.

Again, the model was not able to predict a single metaphor. Adding the emotion regression task did not result in a better performance.

6 Discussion

Our MTL-setup is similar to the ones proposed in (Fornaciari et al., 2021; Dankers et al., 2019). Even though, they showed improved results, we were not able to replicate them for our purpose. Employing auxiliary tasks did not improve the performance. One reason might be, that the number of metaphors in the Kontextbruch is still not enough, even though we increased the number of metaphors through oversampling. The most challenging task is trying to do the most out of the small data set that we have. Increasing the number of auxiliary tasks did not seem to improve the results. Using a on VUA fine-tuned model did not improve the performance as well.

7 Conclusion

In this paper we tried to improve the results of metaphor detection by applying MTL with soft-label regression and emotion regression. All of our models were not able to predict a single metaphor. We were not able to improve the results via weighting the loss functions. With the annotated data, another more related task could have been trying to predict for each token, if it is the focus, frame or none of them in the sentence. Nonetheless, we assume that the data set is too small for the auxiliary tasks to support the main task. The data set has to either be enlarged by adding additional annotated sentences or additional methods have to be found to increase the number of metaphors in the data set.

Appendix

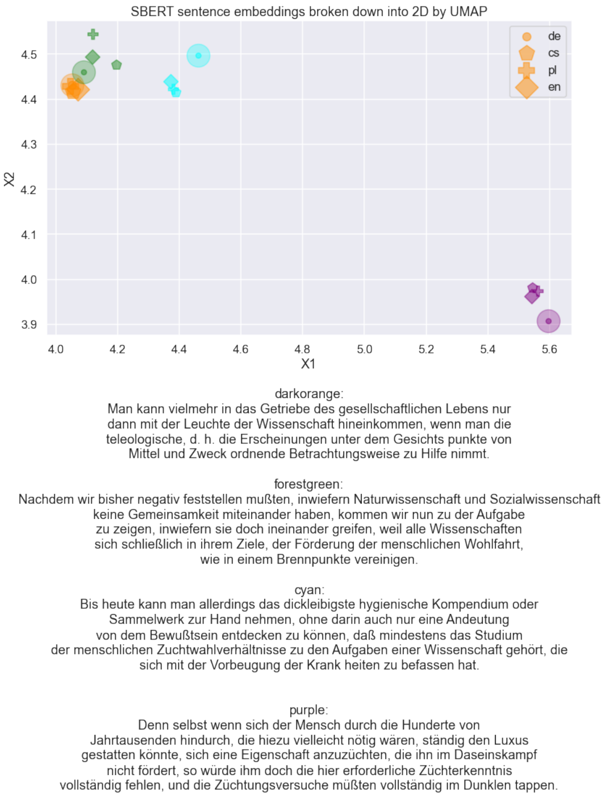
7.1 Work distribution

7.1.1 Fynn Gilbert 3093453

- data section
- Methods:
  - Data augmentation: Oversampling by translation
  - Performance metric: Average precision over F1
  - Loss function: Focal loss over cross entropy
- Experiments + Results:
  - Repeating experiments from previous group
- Github Repository

7.1.2 Dennis Holzmann

- Abstract
- Introduction
- Methods:
  - Chapter 4.2
- Experiments + Results:
  - Chapter 5.2. MTL Experiments with BERT and MBERT
- https://github.com/dholzmann/Metaphor-detection



darkorange:
Man kann vielmehr in das Getriebe des gesellschaftlichen Lebens nur dann mit der Leuchte der Wissenschaft hineinkommen, wenn man die teleologische, d. h. die Erscheinungen unter dem Gesichts punkte von Mittel und Zweck ordnende Betrachtungsweise zu Hilfe nimmt.

forestgreen:
Nachdem wir bisher negativ feststellen mußten, inwiefern Naturwissenschaft und Sozialwissenschaft keine Gemeinsamkeit miteinander haben, kommen wir nun zu der Aufgabe zu zeigen, inwiefern sie doch ineinander greifen, weil alle Wissenschaften sich schließlich in ihrem Ziele, der Förderung der menschlichen Wohlfahrt, wie in einem Brennpunkte vereinigen.

cyan:
Bis heute kann man allerdings das dickleibigste hygienische Kompendium oder Sammelwerk zur Hand nehmen, ohne darin auch nur eine Andeutung von dem Bewußtsein entdecken zu können, daß mindestens das Studium der menschlichen Zuchtwahlverhältnisse zu den Aufgaben einer Wissenschaft gehört, die sich mit der Vorbeugung der Krank heiten zu befassen hat.

purple:
Denn selbst wenn sich der Mensch durch die Hunderte von Jahrtausenden hindurch, die hiezu vielleicht nötig wären, ständig den Luxus gestatten könnte, sich eine Eigenschaft anzuzüchten, die ihn im Daseinskampf nicht fördert, so würde ihm doch die hier erforderliche Züchterkenntnis vollständig fehlen, und die Züchtungsversuche müßten vollständig im Dunklen tappen.

Figure 7: Similar sentences in 2D vectorspace

References

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 41–48, New York, NY, USA. Association for Computing Machinery.

Sven Buechel and Udo Hahn. 2017. EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585, Valencia, Spain. Association for Computational Linguistics.

Rich Caruana. 2004. Multitask learning. *Machine Learning*, 28:41–75.

L. Elizabeth Crawford. 2009. Conceptual metaphors of affect. *Emotion Review*, 1(2):129–139.

Verna Dankers, Marek Rei, Martha Lewis, and Ekaterina Shutova. 2019. Modelling the interplay of metaphor and emotion through multitask learning. 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)*, pages 2218–2229, Hong Kong, China. 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.

Lynn Fainsilber and Andrew Ortony. 1987. Metaphorical uses of language in the expression of emotions. *Metaphor and Symbolic Activity*, 2(4):239–250.

Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, and Massimo Poesio. 2021. Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2591–2597, Online. Association for Computational Linguistics.

Petra Gehring. 2010. *Erkenntnis durch Metaphern? Methodologische Bemerkungen zur Metaphernforschung*, pages 203–220. VS Verlag für Sozialwissenschaften, Wiesbaden.

George Lakoff. 1980. Conceptual metaphor in everyday language. page 77(8).

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection.

Saif Mohammad, Ekaterina Shutova, and Peter Turney. 2016. Metaphor as a medium for emotion: An empirical study. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 23–33, Berlin, Germany. Association for Computational Linguistics.

Duc Dat Nguey, Janosch van Kann, and Matthias Lang. 2021. Deep learning for metaphor detection using the kontextbruch approach.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Gerard Steen, Lettie Dorst, J. Herrmann, Anna Kaal, Tina Krennmayr, and Trijntje Pasma. 2010. *A method for linguistic metaphor identification: From MIP to MIPVU*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.