ITF-NLP at SemEval-2025 Task 11: An Exploration of English and German Multi-label Emotion Detection using Fine-tuned Transformer Models

Samantha Kent and **Theresa Nindel**
Fraunhofer FKIE
Fraunhoferstraße 20, 53343 Wachtberg, Germany
{samantha.kent,theresa.nindel}@fkie.fraunhofer.de

Abstract

We present our submission to Task 11, Bridging the Gap in Text-Based Emotion Detection, of the 19th International Workshop on Semantic Evaluation (SemEval) 2025. We participated in track A, multi-label emotion detection, in both German and English. Our approach is based on fine-tuning transformer models for each language, and our models achieve a macro F1 of 0.75 and 0.62 for English and German respectively. Furthermore, we analyze the data available for training to gain insight into the model predictions.

1 Introduction

The American Psychological Association defines emotion as "conscious mental reactions (such as anger or fear) subjectively experienced as strong feelings usually directed toward a specific object and typically accompanied by physiological and behavioral changes in the body"[1].

Most research in emotion detection in Natural Language Processing (NLP) is based on two different types of emotion theories. The first group of theories views emotions as universal categories (e.g.(Ekman, 1992) and (Plutchik, 1980)), and the second focuses on defining emotions in two or three dimensions, such as valance, arousal and dominance e.g. (Russell, 1980). Furthermore, current research is exploring the possibility of multiple emotions being present in one sentence or utterance. Task 11, Bridging the Gap in Text-Based Emotion Detection, at the International Workshop on Semantic Evaluation (SemEval), focuses on cross-lingual and multi-label emotion detection (Muhammad et al., 2025b).

We participated in Task A 'Multi-label Emotion Detection' for the languages German and English. The aim of the task was to detect the emotions Anger, Fear, Joy, Sadness, and Surprise for English, and the German data also includes the class Disgust. Notably, a text can contain multiple emotions, thus takes into account the co-existence or even potential overlapping of emotions in one sentence. An example in English can be found below:

```
{text: It could have been my eye
   -- but my glasses probably
   blocked that from happening,
   and diverted the injury higher
    up on my head.,
 gold labels: fear and surprise}
```

Previous shared tasks on emotions demonstrate a range of different approaches. Affect in Tweets at Semeval 2018 includes emotion classification systems based on SVMs and LSTMs (Mohammad et al., 2018). Neural architectures were also the most common approach for an emotion shared task related to context in emotion detection (Chatterjee et al., 2019). For the WASSA 2022 shared task, emotion label prediction was conducted for a series of essays using Ekman's six emotion classes. Most teams used systems with pre-trained Transformer mechanisms such as BERT, RoBERTa and DeBERTa (Barriere et al., 2022). Following this, most participants in the WASSA EXALT Shared Task on explainability for Cross-Lingual Emotions in Tweets use some form of Generative Large Language Model (LLM) (Maladry et al., 2024).

Based on the previous approaches to emotion shared tasks, we decide to focus on fine-tuning a transformers model for each language, English and German, and to use the results as a starting point to analyze the emotion labels in more detail. We chose to draw upon well-established discriminative transformer models instead of generative LLMs as our main focus is on advancing the understanding of decisions made by those commonly used baseline models. In general, we are interested in exploring the emotion classes, linking model performance to the data, and comparing the differences between

[1] https://www.apa.org/topics/emotions.

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1465–1472

the two languages.

2 Data

The following chapter describes the data available for training in English and German that was provided by the task organizers. For both languages, the train/development/test instances stem from Reddit and was annotated using language-specific human-annotators. A more detailed overview of the dataset, that contains human-annotated emotion data for 28 different languages, can be found in (Muhammad et al., 2025a).

The **German** data consists of 2803 annotated texts and was labeled for six different emotions: Anger, Disgust, Fear, Joy, Sadness, and Surprise. Each text contains one or more labels, with an average of 1.17 emotion label per text, where 24.76% ($n = 694$) were labeled as not containing any emotions (neutral). 41.42% ($n = 1161$) were labeled with one emotion, 25.76% ($n = 722$) with two, 7.49% ($n = 210$) with three, 0.54% ($n = 15$) with four and 0.04% ($n = 1$) with five different emotions.

The **English** data contains 2884 annotated texts and was labeled for five different emotions: Anger, Fear, Joy, Sadness, and Surprise. 8.74% ($n = 252$) texts were labeled as not containing any emotions (neutral). 41.16% ($n = 1187$) were labeled with one emotion, 37.21% ($n = 1073$) with two, 10.82% ($n = 312$) with three, 2.01% ($n = 58$) with four and 0.07% ($n = 2$) with five different emotions.

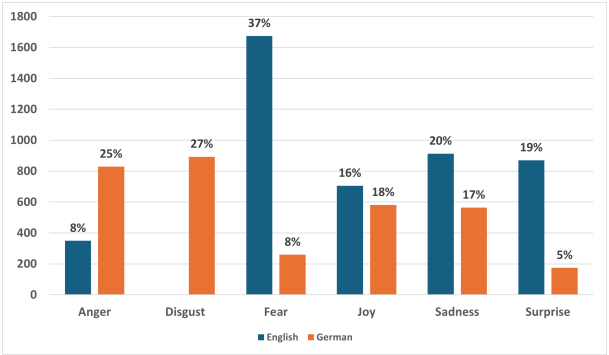

Figure 1: Emotion class distribution for English and German data.

Figure 1 illustrates the class imbalance in both the English and German train datasets. For English, the largest class is Fear (37%), and the smallest is Anger (8%). Contrastingly, in the German data, Fear (8%) is one of the smallest classes and Anger (25%) is the largest.

3 System Description

During the development phase, a number of pretrained models obtained from Hugging Face's model repository[2] were tested with regard to their ability to solve the given task. Model training/evaluation was implemented with the help of the Simple Transformers library (Rajapakse, 2019) and Pytorch (Paszke et al., 2019).

We used a similar approach for both languages and tested various model combinations and train/development/test splits. In submission 1 in both languages, the models were fine-tuned on the training data provided by the organizers. The development data was used to evaluate the models. For the second and third submissions, the training and development data was reshuffled and split into a new a training, development, and held out test set (distribution 70/20/10%). Additionally, weights were calculated for all classes and used during training to lessen the effects of the uneven class distribution for each language individually.

The **English** models submitted during the test phase were based on DeBERTa (He et al., 2020) and RoBERTa (Liu et al., 2019). The DeBERTa model, used for submission 1, was trained using the training data provided by the organizers for 5 epochs with the following hyperparameters: training batch size: 32, learning rate: 2e-5, max length: 125. For submissions 2 and 3, a RoBERTa model was fine-tuned using an 80/20 train and development split. The parameters are similar to the previous model, except training was conducted with a learning rate of 3e-5 and max length was set to 100. The difference between the two models in submission 2 and 3 is based on a different train split, providing the models with different data for fine-tuning.

The best performing model for **German**, which was subsequently used for all submissions, was xlm-roberta-large-finetuned-conll03-german (Conneau et al., 2019). As the name suggests, the model is based on XLM-RoBERTa-large (Conneau et al., 2019) and was fine-tuned on a German dataset. Hyperparameter testing resulted in the following optimal parameter combination: Epochs: 5, Learning rate: 3e-5, Training batch size: 16.

4 Results

Table 1 below shows results of the submissions that were made using the final test data provided by the

[2]https://huggingface.co/models

organizers. They are similar to the monolingual results reported by the organizers (Muhammad et al., 2025a).

For **English**, the best performing model achieved a macro F1 of 0.7501 (submission 3), and is also the submission on the final rank list. Submission 1, the DeBERTa model, did not perform as well as the other two RoBERTa models. This is surprising considering it outperformed the other two models on our own test data. There was a much larger drop in performance between our own test set and the final set for this model compared to the two RoBERTa models. Interestingly, the model for submission 1 was trained without using weighting to balance out the uneven class distribution, as the performance on our own test set was similar with or without weighting. The weights were used for model training in submissions 2 and 3. This suggests that using weighing as a strategy to balance the classes contributed to the robustness of the models.

	1	2	3
English			
Own test data	0.783	0.752	0.774
Final test data	0.6991	0.7485	**0.7501**
German			
Own test data	0.6608	0.6445	0.6414
Final test data	0.6231	0.6131	**0.6222**

Table 1: Results for English and German on final test data. The best performing models from the official ranking are in bold.

For **German**, the macro F1 scores ranged from 0.6131 to 0.6231. A drop in macro F1-scores between the self-compiled test set and the test set provided by the organizers can be observed for all three submissions. The smallest difference is present in the scores achieved by model 3. While model 3 has the lowest macro F1-score on the self-compiled test set, it seems to be the most robust when it comes to the prediction of previously unseen data. Further analysis would be needed to evaluate if the greater difference between macro F1-scores for the first two submissions could be due to over-fitting.

As well as looking at the overall performance, we also inspected the model's performance on the individual classes. The **German** models struggled to correctly predict the classes Fear and Surprise specifically. We showed in the data description that these classes are underrepresented in the German data. The weighting that was implemented during fine-tuning to lessen the effect of the unbalanced data distribution was not sufficient. The best performing classes, Anger and Joy, are also highly represented in the data.

	English	**German**
Anger	0.6621	0.7536
Disgust	-	0.6987
Fear	0.8398	0.4784
Joy	0.7546	0.7389
Sadness	0.7621	0.6443
Surprise	0.7316	0.4192
Macro F1	0.7501	0.6222

Table 2: Fine-grained results on submission 3 for English and German.

For **English**, a similar pattern of emotion class size and the model's ability to accurately predict the class can be observed. Fear outperforms the other classes with an F1 of 0.8398, and Anger is by far the smallest emotion class and also achieves the lowest f-score.

5 Analysis

In this section we analyze the corpus data and link the results to possible performance issues in the models. We further explore specific emotion classes, namely Anger and Disgust in German and Fear and Sadness in English.

5.1 German

To gain more insights into the performance of our best German model (model 3), the data available for training as well as the errors made by the model were analyzed. Out of the 2604 German sentences in the test data, only 1274 (48.92%) are correct, with all possible labels correctly predicted by the model. There is a difference when analyzing the per class or per sentence predictions. Even though the model achieves a macro-F1 of 0.62, when taking into account whether all emotions are predicted correctly, the results are not as accurate. Figure 2 below shows the distribution of the number of labels per sentence for all sentences that were predicted incorrectly. About 80% of errors stem from sentences that should contain one or two labels, with the majority of misclassifications being due to the models predicting multiple emotions where only one is correct. Therefore, in a first analy-

sis step, we want to evaluate the most frequent label combinations in the training data to determine whether overlapping or co-occurring emotions are the source of some of the German model errors.

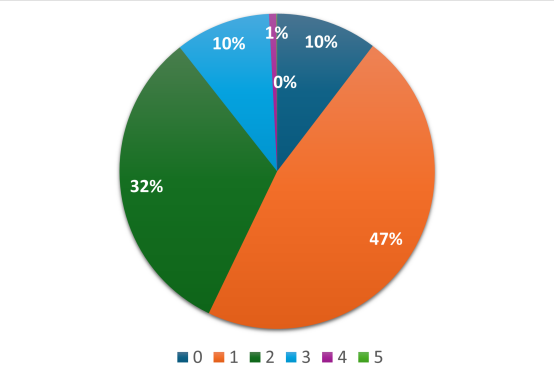

Figure 2: Number of labels per wrong prediction (de)

In total, the ten emotion combinations in figure 3 account for 2417 instances of all available data points (86.23% of the train/development corpus). Notably, the emotions Anger and Disgust appear together in two of these top 10 combinations and make up the biggest two-label combination with 384 (13.70%) entries in the corpus. Therefore, to ascertain how similar the classes are, we decided to analyze the most common words included in the texts associated with the two emotion categories in more detail. Stop words were not considered for this analysis.

Label(s)	**Nr.**
No Emotions/Neutral	694
Joy	435
Anger, Disgust	384
Sadness	220
Disgust	181
Anger	158
Anger, Disgust, Sadness	118
Fear	106
Surprise	61
Disgust, Sadness	60

Table 3: Top 10 Label (Combinations) in the German Dataset

The results in figure 4 in appendix A show a high lexical overlap between the categories Anger and Disgust. Nouns such as 'Israel', 'Krieg' (war), 'Gaza' and 'Hamas' are frequent in both emotions. As the emotions often co-occur in the data, this is not surprising. However, the question arises if this co-occurrence leads to a tendency of the model to learn similar representations based on the content of the classes rather than the associated emotions. To examine this question, we explore sentences that were annotated as only Anger, but have instead been classified as Disgust or a combination of Anger and Disgust, and vice versa.

Gold Labels	**Predictions**	**%**
Anger (140)	Disgust	61 (43.57)
	Fear	6 (4.29)
	Joy	6 (4.29)
	Sadness	1 (0.71)
	Surprise	7 (5)
Disgust (164)	Anger	46 (28.05)
	Fear	6 (3.66)
	Joy	8 (4.88)
	Sadness	17 (10.37)
	Surprise	6 (3.66)

Table 4: Classification Errors for Anger and Disgust.

Table 4 illustrates that our best German model predicted either Disgust or a combination of Anger and Disgust instead of the correct label Anger, in 43.57% of all errors related to Anger (as a single emotion annotation). Contrastingly, Joy, for example, was only predicted in combination with Anger in 4.29% of cases. This is also reflected in our train data: Joy and Anger, as well as Fear and Anger, are only labeled in combination 14 times, whereas Anger and Disgust are present as a label combination in 384 sentences (see table 3).

Even though there may also be other reasons for this type of misclassification, the very similar vocabulary in both the Anger and Disgust classes is likely to cause difficulties. The following example serves to illustrate this mix-up. Here, the relevant words are *Krieg* and *Ukraine*. Both are among the most frequent words in both categories, Anger and Disgust, in the data available for training.

```
{text: Einfache Wahrheiten: Wer "
   gegen Krieg ist", sollte die
   Ukraine bestmöglich bei ihrer
   Verteidigung unterstützen. Wer
    diese Unterstützung ablehnt,
   unterstützt de facto Putin in
   seinem Krieg.
 gold label: anger
 predicted labels: anger and
    disgust}
```

In general, our analysis seems to indicate that

the model may in fact be learning similar representations for the emotions Anger and Disgust. This is understandable considering the distribution of classes in the data, as well as the analysis regarding the most common words. The performance of the model in both classes is good, but nonetheless, the multi-label aspect of the task means it might be difficult to actually distinguish between these two classes.

5.2 English

We adopt the same method of analysis for the English dataset In order to determine whether a similar pattern is present in English. With a total of 2884 sentences in the train data, the top 10 label combinations account for 85.64% of the dataset. Similarly to German, we see one two-label combination that occurs frequently in the training data, namely, Fear and Sadness. There does seem to be such a strong co-occurrence of only two specific labels as in German, because Fear and Surprise also often occur in the same sentence. To ensure comparability between the two languages we again further analyze the sentences containing the top two co-occurring emotions.

To start with, the most common words of these two emotion categories were analyzed. The results can be found in appendix A, figure 5. Nouns such as 'head', 'eyes', 'hand' and 'heart' stand out at a first glance, indicating that there does seem to be a common topic in both emotions. However, there are also many verbs present in the top words and the difference in frequency of occurrence is a bit larger between the two emotions compared to in German. Whilst a similar pattern can be observed for English, there does not seem to be such a strong indication of overlapping topics in English Fear and Sadness compared to the German Anger and Disgust.

A more in-depth analysis of the the test data shows that a total of 1381 English sentences (49.91%), out of the possible 2767 sentences in the English test data, are incorrectly classified by the RoBERTa model. Figure 3 illustrates that more than 50% of errors stem from sentences that contain more than one emotion prediction, but also for English a large percentage of errors is due to misclassifications in sentences containing single emotions.

Based on the frequent label combinations in the training data, we further explore if the class combinations learned during training also influence the predicted labels. Table 6 provides an overview of misclassifications related to the most frequent emotion combination Fear and Sadness. As is true for the German analysis, the figures in the table show the relationship between the misclassifications in those emotion classes, but do not account for all possible combinations. Also as expected based on the German data, classes that frequently appear together in the train data, also seem to be a source of error for the predictions.

Label(s)	**Nr.**
Joy	448
Fear, Sadness	429
Fear	425
Fear, Surprise	337
No Emotions/Neutral	252
Sadness	139
Fear, Sadness, Surprise	127
Surprise	117
Joy, Surprise	114
Anger, Fear, Sadness	82

Table 5: Top 10 Label (Combinations) in the English Dataset

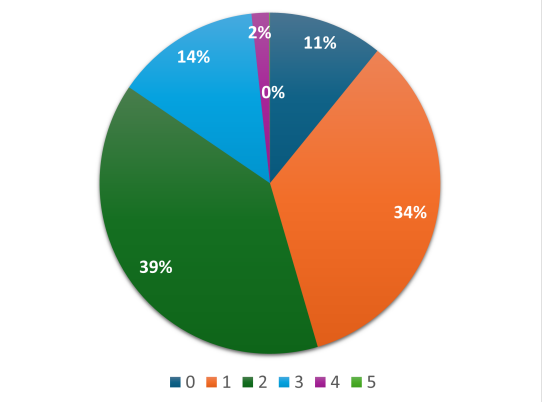

Figure 3: Number of labels per wrong prediction (en)

Gold Labels	**Predictions**	**%**
Fear (172)	Sadness	56 (32.56)
	Anger	7 (4.07)
	Joy	14 (8.14)
	Surprise	45 (26.16)
Sadness (70)	Fear	41 (58.57)
	Anger	4 (5.71)
	Joy	9 (12.86)
	Surprise	3 (4.29)

Table 6: Classification Errors for Fear and Sadness.

In the example sentence below, the label should have been predicted as Fear, but Sadness has also been included as a prediction label. The most frequent words from the classes Fear and Sadness in figure 5 in appendix A, show that *head* is frequent in both classes, but more dominant in Fear.

```
{text: Meanwhile my head began to
    pound and I grew quite
   nauseous.
 gold label: fear
 predicted labels: fear and
    sadness}
```

6 Conclusion

We participated in the Semeval 2025 shared task on text-based emotion detection. We participated in task A in English and German and our results are a macro F1 of 0.75 and 0.62 respectively. These are similar to those achieved by the organizers for their monolingual models (Muhammad et al., 2025a). Our approach was based on fine-tuning well-established transformers models and we optimized the model parameters for each language and experimented with different approaches to optimizing the training data. We found that the most effective strategy to improve both performance and robustness in the models for both languages was to balance the emotion classes in the data.

In general, our analysis of the train and test data suggests, for both German and English, that there is a connection between class size and model performance for that specific class. We also demonstrated with our analysis the difficulty in classifying more closely related emotions, specifically Anger and Disgust in German, and Fear and Sadness in English. This seems to be related to their more frequent co-occurrence as annotated labels in the training data, which then also has an effect on how closely related the topics in each class are. Future work includes expanding the emotion correlation analyses and applying our findings when balancing the dataset in pre-processing.

When comparing the two languages, based on the most frequent words it seems as though a majority of the sentences in German data are related to politics, whereas in the English data the prevailing topic seems to be health. A larger dataset with more diverse topics might be helpful in ensuring robustness in future models. It would be interesting to explore the topics present in the data for the other languages in the shared task, and also see what role the topic clusters may play in cross-lingual emotion detection.

In general, the definition of emotion already suggests that subjectivity plays a large role in correctly perceiving emotion, and multi-label annotations make the task of emotion detection even more challenging. The example below serves to illustrate the need for annotating multiple emotions in one sentence, as there is evidently an expression of both negative and positive emotion. However, due to the subjectivity of perceiving emotion, the need for all five emotion labels is debatable. We therefore acknowledge the difficulty of collecting and annotating emotion data.

```
{text: Yeah...welcome to being 25
   , btw...it is awful thus far
   ...but...SHIT at least I get
   to be 25!
 gold label: anger, fear, joy,
    sadness, surprise}
```

References

Valentin Barriere, Shabnam Tafreshi, João Sedoc, and Sawsan Alqahtani. 2022. WASSA 2022 shared task: Predicting empathy, emotion and personality in reaction to news stories. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 214–227, Dublin, Ireland. Association for Computational Linguistics.

Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi, and Puneet Agrawal. 2019. SemEval-2019 task 3: EmoContext contextual emotion detection in text. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 39–48, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced BERT with disentangled attention. *CoRR*, abs/2006.03654.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis,

Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. Cite arxiv:1907.11692.

Aaron Maladry, Pranaydeep Singh, and Els Lefever. 2024. Findings of the WASSA 2024 EXALT shared task on explainability for cross-lingual emotion in tweets. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 454–463, Bangkok, Thailand. Association for Computational Linguistics.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine de Kock, Nirmal Surange, Daniela Teodorescu, Ibrahim Said Ahmad, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino D. M. A. Ali, Ilseyar Alimova, Vladimir Araujo, Nikolay Babakov, Naomi Baes, Ana-Maria Bucur, Andiswa Bukula, Guanqun Cao, Rodrigo Tufino Cardenas, Rendi Chevi, Chiamaka Ijeoma Chukwuneke, Alexandra Ciobotaru, Daryna Dementieva, Murja Sani Gadanya, Robert Geislinger, Bela Gipp, Oumaima Hourrane, Oana Ignat, Falalu Ibrahim Lawan, Rooweither Mabuya, Rahmad Mahendra, Vukosi Marivate, Andrew Piper, Alexander Panchenko, Charles Henrique Porto Ferreira, Vitaly Protasov, Samuel Rutunda, Manish Shrivastava, Aura Cristina Udrea, Lilian Diana Awuor Wanzare, Sophie Wu, Florian Valentin Wunderlich, Hanif Muhammad Zhafran, Tianhui Zhang, Yi Zhou, and Saif M. Mohammad. 2025a. Brighter: Bridging the gap in human-annotated textual emotion recognition datasets for 28 languages. *Preprint*, arXiv:2502.11926.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Seid Muhie Yimam, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine De Kock, Tadesse Destaw Belay, Ibrahim Said Ahmad, Nirmal Surange, Daniela Teodorescu, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino Ali, Vladimir Araujo, Abinew Ali Ayele, Oana Ignat, Alexander Panchenko, Yi Zhou, and Saif M. Mohammad. 2025b. SemEval task 11: Bridging the gap in text-based emotion detection. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, Vienna, Austria. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. *CoRR*, abs/1912.01703.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. *Theories of emotion*, 1:3–31.

T. C. Rajapakse. 2019. Simple transformers. `https://github.com/ThilinaRajapakse/simpletransformers`.

James A. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178.

A Appendix

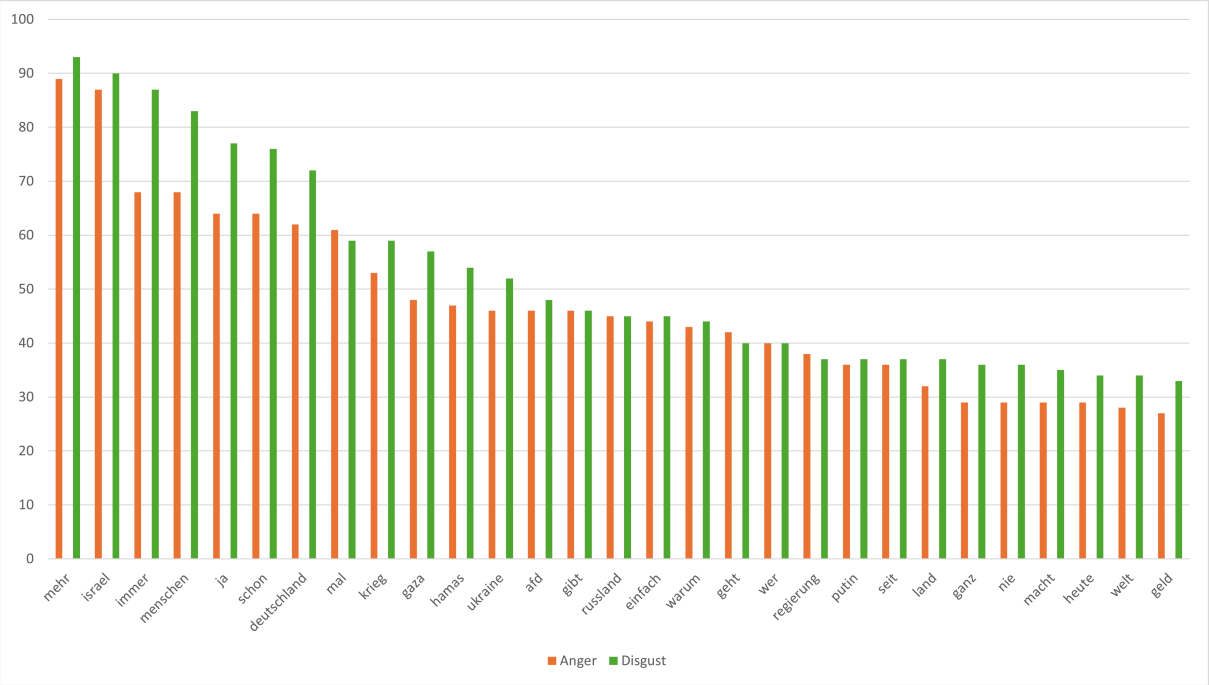

Figure 4: Top Words for Anger and Disgust in German

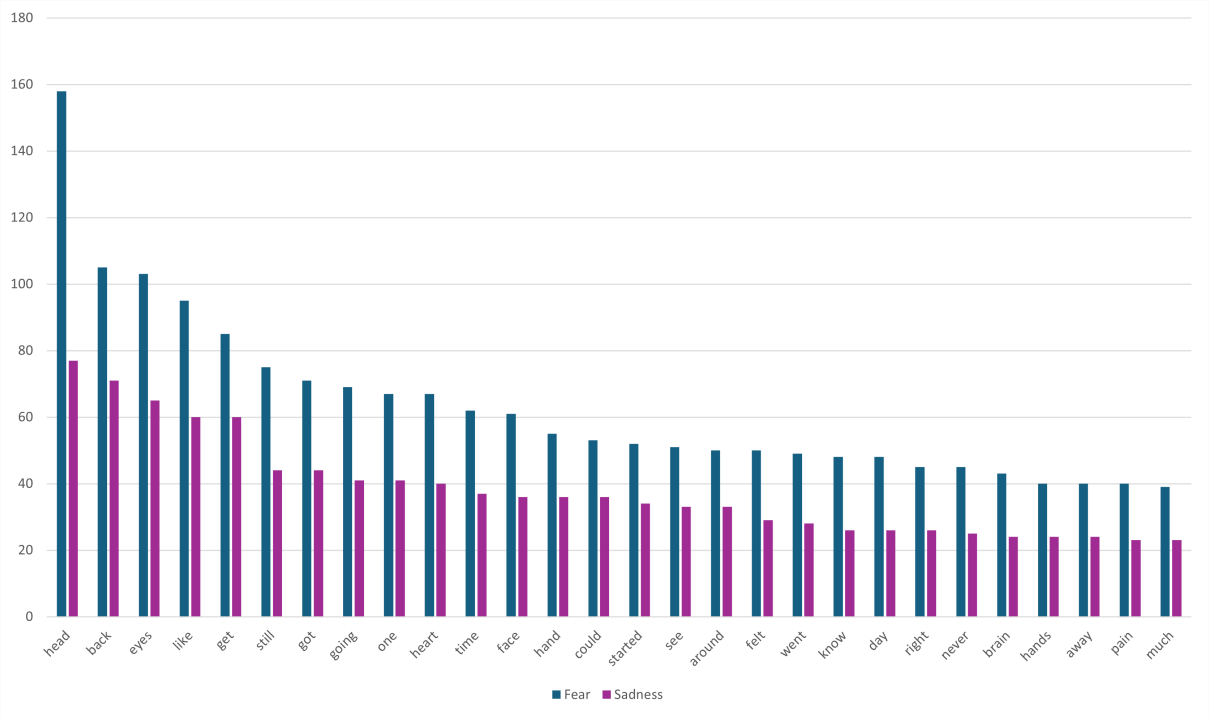

Figure 5: Top Words for Fear and Sadness in English