Team KiAmSo at SemEval-2025 Task 11: A Comparison of Classification Models for Multi-label Emotion Detection

Kimberly Sharp, Sofia Kathmann and **Amelie Rüeck**
University of Tübingen
Department of General and Computational Linguistics
{ kimberly.sharp,sofia.kathmann,amelie.rueeck } @student.uni-tuebingen.de

Abstract

The aim of this paper is to take on the challenge of multi-label emotion detection for a variety of languages as part of Track A in SemEval 2025 Task 11: Bridging the Gap in Text-Based Emotion Detection. We fine-tune different pre-trained mono- and multilingual language models and compare their performance on multi-label emotion detection on a variety of high-resource and low-resource languages. Overall, we find that monolingual models tend to perform better, but for low-resource languages that do not have state-of-the-art pre-trained language models, multilingual models can achieve comparable results.

1 Introduction

Interlocutors rarely speak in an entirely neutral manner: more often than not, speakers will use emotions in their speech. Emotions are an important driving force of conversations and understanding how language and emotions interact is crucial for linguistics. In NLP, the field of *Emotion Recognition* is concerned with identifying the emotions of the speaker of an utterance. Ekman (1992) defines six basic emotional states: joy, sadness, fear, anger, surprise, and disgust. These have since been used widely in emotion recognition to assign the perceived emotion of a speaker during an utterance.

While many systems have been developed that can assign one singular emotion to a text, the challenge of *Multi-label Emotion Detection* is a newer and less investigated field. Nonetheless, it is an important area, since speakers rarely feel emotions in isolation and multiple emotions often occur in conjunction with each other, for example anger and disgust.

Track A of SemEval 2025 Task 11, "Bridging the Gap in Text-Based Emotion Detection" (Muhammad et al., 2025b), aims at solving the issue of multi-label emotion detection for 28 different languages (Muhammad et al., 2025a), including many low-resource languages. To solve this task, our team compares different pre-trained language models on their ability to perform multi-label emotion classification, comparing monolingual models like GottBERT (Scheible et al., 2024) and Twitter-RoBERTa (Barbieri et al., 2020) to multilingual models such as XLM-T (Loureiro et al., 2022). Our aim is to see how multilingual models perform when they are fine-tuned solely on one language versus multiple languages simultaneously. This could provide important insights into maximizing the usability of multilingual models for low-resource languages. We further employ task-adaptive pre-training and optimized classification thresholds at each epoch to improve performance.

2 Background

Early work in NLP largely focused on *Sentiment Analysis*, the classification of a text into negative or positive valence classes (Mohammad and Kiritchenko, 2018). In contrast, *Emotion Recognition* deals with assigning texts to distinct emotion classes. Sentiment Analysis is often a case of binary classification, assigning either positive or negative valence to a text. Emotion Recognition is often implemented as a multi-class classification problem, selecting the most salient emotion out of multiple emotion classes. However, a multi-class approach neglects the co-occurrence of emotions that cannot be separated from each other (Mohammad and Kiritchenko, 2018). This type of relation requires a multi-label strategy. Furthermore, most existing multi-label emotion classifiers focus on high-resource, predominantly Indo-European languages such as English, with fewer systems available for low-resource languages.

Earlier approaches to multi-label emotion recognition employed *classifier chains* to account for the correlation between the different emotions. For instance, participants of SemEval 2018 Task 1 com-

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1542–1548
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

bined their best performing classifier for every emotion into a chain which passes the predicted labels to the next classifier in the chain, sorting the classifiers by performance (highest to lowest). Their best performing classifier chain achieved a macro-averaged F1 score of 0.493 (De Bruyne et al., 2018).

More recent approaches involve *Latent Emotion Memory* networks, which aim at learning the latent emotion distribution and emotion intensity in a text and leverage it into a classification system. These consist of a variational auto-encoder that learns the emotion from the input and a memory unit that captures the most salient features for that emotion. On the SemEval 2018 dataset, they achieved a macro F1 score of 0.567 (Fei et al., 2020).

Other systems include the *Sequence-to-Emotion (Seq2Emo)* approach, which is essentially a sequence-to-sequence model that encodes the utterance using an LSTM and then uses an LSTM-based decoder to perform binary classification on the emotions sequentially. This approach achieved a macro F1 score of 0.5192 on the SemEval 2018 dataset (Huang et al., 2021).

Since then, the rise of pre-trained language models and transformer-based architectures has opened up a variety of new ways to approach multi-label emotion detection. However, it is still unclear which pre-trained models are well suited for emotion detection, and how to best fine-tune models for this task. A further open question is how to build multilingual models that can perform emotion detection in a variety of languages.

This is the aim of our approach: We compare several monolingual and multilingual pre-trained language models and fine-tune them for emotion classification, comparing the pre-trained models to a logistic regression baseline. The following sections will explore the different systems that we have tried and their respective results.

3 System Overview

For our system, we mainly rely on the XLM-Twitter (XLM-T) base model for sequence classification (Barbieri et al., 2022), which continues pre-training from a publicly available XLM-R checkpoint (Conneau et al., 2020) using nearly 200M tweets from over 30 languages. We then apply different fine-tuning strategies and observe the effects on model performance. Additionally, we contrast the performance of a multilingual model like XLM-T with specialized monolingual models for German and English. Due to time and resource constraints, we only analyze a subset of 10 languages that includes both high-resource and low-resource languages: Afrikaans, Amharic[1], Algerian Arabic, Moroccan Arabic, Mandarin Chinese, German, English, Spanish, Hausa, and Hindi. We use the training, development, and test datasets provided by the SemEval2025 Task 11 organizers (Muhammad et al., 2025a; Belay et al., 2025).

3.1 Linear Baseline

Nowadays, traditional machine-learning algorithms such as Logistic Regression or Random Forests are often overlooked in favour of transformer-based architectures. Nonetheless, their cost-effectiveness and explainability make them an interesting baseline that can provide a useful reference point for evaluating transformer-based approaches.

For our baseline, we convert the input texts into sparse tf-idf vectors and train a Logistic Regression classifier using the One-vs-Rest (OVR) multiclass strategy. This strategy consists of training one binary classifier independently for each label – each classifier fits the current label against all the other labels.

When running our experiments, this simple baseline achieved an F1 score similar to XLM-T for 4 out of 10 languages and even outperformed it for Hausa (see Table 1).

3.2 Fine-tuning monolingual models

To better contextualize the performance of the multilingual XLM-T, we fine-tune a specialized, fully monolingual model for German and English respectively. Due to its well-suitedness to the task data, we chose Twitter-RoBERTa (Loureiro et al., 2022) for English. There were considerably fewer options for German, so we decided on the GottBERT base model (Scheible et al., 2024), which is not pre-trained on tweet data, but is based on the RoBERTa architecture (Liu et al., 2019). We then fine-tune both models on their respective language data for Track A.

3.3 Fine-tuning a multilingual model

In order to be able to run the task of emotion detection on languages with less resources available than German and English, we leverage the pre-trained

[1] Belay et al. (2025) provide the datasets for the Ethiopian languages Amharic, Oromo, Somali, and Tigrinya.

multilingual large language model XLM-T. We first fine-tune the model on the joint training data for all 10 languages in our sample, resulting in a single multilingual classifier for all languages. To compare, we then fine-tune the same model, XLM-T, on the training data for each language, resulting in one classifier per language.

3.4 Task-adaptive pre-training

Researchers like Gururangan et al. (2020), as well as submissions to previous years of SemEval, for example by Wang et al. (2023), have shown the effectiveness of continuing to pre-train and adapt large language models that have so far been trained on huge, heterogeneous corpora. Domain-adaptive and task-adaptive pre-training – continued pre-training with domain- and task-specific data – consistently improves performance on the domains and tasks the additional data is from. Since XLM-T builds on XLM-R by training on tweet data, it can be said to already come with a certain amount of domain-adaptive pre-training off-the-shelf. Additionally, when exploring the effects of language-adaptive pre-training (domain-adaptive pre-training where the target language is considered to be the domain) and task-adaptive pre-training on multilingual sentiment analysis, Wang et al. (2023) find task-adaptive pre-training to be the main contributor to improved classifier performance. Therefore, and also due to time and resource constraints, we only apply a minimal version of task-adaptive pre-training (TAPT).

For that, we continue training our XLM-T model on the original masked language modeling (MLM) training objective, using the unlabeled training data from the 10 languages in our sample. We then fine-tune it for emotion classification on the joint data of 4 languages: German, English, Spanish and Hindi.

3.5 Fine-tuning a T5 model

We also further investigate the T5 pre-trained model. We use the T5 base model (Raffel et al., 2020) initially on English only and then move to T5 fine-tuned for Emotion Recognition (Romero), as it has similar emotion labels to our task. Although both T5 models used are English monolingual models, we run the fine-tuned model on German as well for comparison purposes. In an early analysis, the results for English of the T5 fine-tuned model were competitive to the scores obtained with XLM-T without fine-tuning. However, due to T5 being outperformed by the English Twitter-RoBERTa model, as well as the lack of a T5 model fine-tuned specifically on tweet data, we focus on the RoBERTa-based models. Nonetheless, we believe that T5 achieving similar results to a RoBERTa-based model may be indicative of further research into T5-based models possibly proving successful in a monolingual framework.

4 Experimental Setup

In this section, we provide details on our considerations about the data and training of the models.

4.1 Datasets

Mentions of usernames in the data have already been replaced by "@<username>", and URLs by "##URL##" in the datasets distributed by the task organizers. Since this low-impact, potentially sensitive data has already been cleaned, and to preserve all meaningful features in the data, we do not apply any further preprocessing.

Before training, we combine the training and development sets to make our own stratified training and validation splits using the *skmultilearn* library by Szymański and Kajdanowicz (2017).

4.2 Training with optimized thresholds

The training data contains large class imbalances between the different emotions, making some emotions harder to learn than others. To account for this, we optimize individual thresholds for each emotion to allow for lower thresholds for smaller classes (leading to a higher recall) and higher thresholds for larger classes (leading to a higher precision). At each training epoch, we start with a preliminary threshold of 0.5 for each emotion. After the epoch, we evaluate the current model on a validation set, and then iteratively adjust the thresholds until we reach the best possible macro-averaged F1 score. We then re-run the predictions with the optimized thresholds and calculate the loss. The model with the currently best F1 score is saved as our checkpoint.

To evaluate the model on the development set, we again compute an individual classification threshold for each emotion using the same strategy. Then for running inference on the test set, we directly apply the thresholds of the best training epoch to the classification.

4.3 Training resources

For mono- and multilingual fine-tuning, we use the AdamW optimizer (Loshchilov and Hutter, 2019)

Model	afr	amh	arq	ary	chn	deu	eng	esp	hau	hin
Logistic Regression	0.2029	0.5369	0.4387	0.3062	0.0881	0.4425	0.4912	0.6114	**0.6048**	0.5628
XLM-T										
Monolingual fine-tuning	0.4673	0.5345	0.4912	**0.5013**	0.5664	0.5332	0.6450	**0.7748**	0.5837	**0.8078**
Multilingual fine-tuning	**0.5034**	**0.6073**	**0.5052**	0.4722	0.5862	0.5813	0.6448	0.7672	0.5425	0.7644
XLM-T with TAPT										
Multilingual fine-tuning	0.3927	0.4925	0.4726	0.4280	**0.6644**	0.5773	0.6561	0.7635	0.4075	0.7855
True monolingual models										
GottBERT	–	–	–	–	–	**0.5976**	–	–	–	–
Twitter-RoBERTa	–	–	–	–	–	–	**0.7251**	–	–	–
T5 model										
Base	–	–	–	–	–	–	0.5939	–	–	–
Fine-tuned	–	–	–	–	–	0.4536	0.6541	–	–	–
SemEval Baseline	0.3714	0.6383	0.4141	0.4716	0.5308	0.6423	0.7083	0.7744	0.5955	0.8551

Table 1: Overview of macro-averaged F1-scores for all our models and analyzed languages

with an initial learning rate of 1e-5 and a maximum number of 10 epochs. For task-adaptive pre-training, we use AdamW with a learning rate of 5e-5. We were only able to run TAPT for 3 epochs. For both fine-tuning and continued pre-training we use a batch size of 16 and a maximum sequence length of 150.

All transformer-based architectures were trained on T4 or L4 GPUs as available through Google Colab and relying on the *Huggingface Transformers* library (Wolf et al., 2020). The Logistic Regression classifier was trained using the *sklearn* library (Pedregosa et al., 2011).

5 Results

Overall, we were able to outperform the SemEval Baseline in 7 out of 10 submitted languages, only for Amharic, German, and Hindi we were not able to achieve a score above the baseline. Table 1 shows the results from all our experiments, while Table 2 shows our final submission results. Since we ran some of the experiments after the end of the evaluation phase, we were not able to submit our final best scores for all languages. XLM-T achieves the best results in 7 out of 10 languages, although some languages benefit more from monolingual fine-tuning, while others do better with multilingual fine-tuning. For German and English, their specialized monolingual models GottBERT and Twitter-RoBERTa outperform XLM-T, regardless of the fine-tuning strategy. Interestingly, the best performing model for Chinese is XLM-T with task-adaptive pre-training (TAPT) and joint fine-tuning on four languages, even though Chinese had not been in the set of languages that model was fine-tuned on.

With the exception of Chinese, we could not replicate previous findings showing that applying task-adaptive pre-training significantly increases model performance. In fact, its performance for Afrikaans and Hausa is quite weak in comparison. However, this system still achieves competitive results for the majority of the languages. We suppose that due to our limited resources, we were not able to fully tap into the potential of TAPT, as the pre-training process was aborted after 3 epochs, which is not nearly enough time for the model to converge. This exactly might have been the issue with Afrikaans and Hausa, which are also the only two languages in our set not present in the top 30 languages XLM-T was originally trained on (Barbieri et al., 2022).

Our resource limitations for applying task-adaptive pre-training are an example for the trade-off between performance and resource use that researchers in this field are continuously faced with. On that note, it is interesting to remark that in our experiments, Logistic Regression slightly outperforms XLM-T for Hausa. We do not have a solid hypothesis for this, especially since calculating the SCUMBLE score (Charte et al., 2019) for our 10 languages suggests that Hausa, along with Chinese, has the highest label concurrency (minority labels occurring mostly or only together with majority labels), which should make it especially difficult to get accurate classification results for their minority labels.

When comparing the performance of joint multilingual and monolingual fine-tuning, there seems to be no clear winner at first. Taking into account whether the target language is low-resource or high-resource however, there seems to be a tendency for low-resource languages to prefer multilingual training. Afrikaans, and especially Amharic, seem to benefit from the additional information present in the other language data with an increase of the F1 score from 0.53 to 0.6 for the latter. Conversely, high-resource languages mostly perform better with monolingual training. This is especially highlighted when comparing the language-specific models GottBERT and Twitter-RoBERTa with the multilingual XLM-T fine-tuned on German or English data. Both fully monolingual models outperform the multilingual one.

For completeness, it would be interesting to compare the performance of XLM-T with TAPT, fine-tuned for each language data individually, with our jointly fine-tuned TAPT-applied XLM-T. It remains an open question whether with our setup we would reach a similar conclusion as Wang et al. (2023), where the advantages of monolingual training become less pronounced in the presence of task-adaptive pre-training.

Language	Micro F1	Macro F1
Afrikaans	0.5236	0.4673
Amharic	0.5566	0.5345
Arabic (Algerian)	0.5118	0.4912
Arabic (Moroccan)	0.5111	0.5013
Chinese	0.6902	0.5664
German	0.6537	0.5976
English	0.7537	0.7251
Spanish	0.7338	0.7635
Hausa	0.5887	0.5837
Hindi	0.7762	0.7855

Table 2: Submission scores for our languages

With our submitted results, we ranked 18th for Afrikaans, 20th for Moroccan Arabic, 21st for Algerian Arabic, 22nd for Hausa, 23rd for Spanish, 24th for German, 26th for Amharic, 26th for Chinese, 31st for Hindi, and 37th for English in the final ranking.

6 Conclusion

Overall, our systems aimed at comparing the performance of mono- and multilingual pre-trained language models for multi-label emotion recognition. We find that when the necessary resources are available, a specialized monolingual approach outperforms a generalized multilingual one. Emotion recognition for high resource languages like German and English works best without the interference of other languages.

Nonetheless, the strength of multilingual models lies in their versatility and their ability to leverage information from higher-resource languages to make inferences about lower-resource languages. As such, multilingual models allow us to tackle tasks with low-resource languages where a specialized approach is simply not feasible. Our example of Chinese shows that classifiers can strongly benefit from being fine-tuned on a set of languages they are not even a part of. Identifying those source languages that are especially useful for improving classification performance in a target language is a task that researchers tackle in the field of zero-shot classification (Lin et al., 2019), which was also the focus of Track C in this SemEval task.

References

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266, Marseille, France. European Language Resources Association.

Tadesse Destaw Belay, Israel Abebe Azime, Abinew Ali Ayele, Grigori Sidorov, Dietrich Klakow, Philip Slusallek, Olga Kolesnikova, and Seid Muhie Yimam. 2025. Evaluating the capabilities of large language models for multi-label emotion understanding. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3523–3540, Abu Dhabi, UAE. Association for Computational Linguistics.

Francisco Charte, Antonio J. Rivera, María J. del Jesus, and Francisco Herrera. 2019. Dealing with difficult minority labels in imbalanced mutilabel data sets. *Neurocomputing*, 326-327:39–53.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Asso-*

ciation for Computational Linguistics, pages 8440–8451, Online. Association for Computational Linguistics.

Luna De Bruyne, Orphée De Clercq, and Véronique Hoste. 2018. LT3 at SemEval-2018 task 1: A classifier chain to detect emotions in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 123–127, New Orleans, Louisiana. Association for Computational Linguistics.

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

Hao Fei, Yue Zhang, Yafeng Ren, and Donghong Ji. 2020. Latent emotion memory for multi-label emotion classification. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7692–7699. AAAI Press.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Chenyang Huang, Amine Trabelsi, Xuebin Qin, Nawshad Farruque, Lili Mou, and Osmar Zaïane. 2021. Seq2Emo: A sequence to multi-label emotion classification model. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4717–4724, Online. Association for Computational Linguistics.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing Transfer Languages for Cross-Lingual Learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

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. *Preprint*, arXiv:1907.11692.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *Preprint*, arXiv:1711.05101.

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-collados. 2022. TimeLMs: Diachronic language models from Twitter. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 251–260, Dublin, Ireland. Association for Computational Linguistics.

Saif Mohammad and Svetlana Kiritchenko. 2018. Understanding emotions: A dataset of tweets to study interactions between affect categories. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Manuel Romero. T5-base fine-tuned for emotion recognition. https://huggingface.co/mrm8488/t5-base-finetuned-emotion.

Raphael Scheible, Johann Frei, Fabian Thomczyk, Henry He, Patric Tippmann, Jochen Knaus, Victor Jaravine, Frank Kramer, and Martin Boeker. 2024. GottBERT: a pure German language model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21237–21250, Miami, Florida, USA. Association for Computational Linguistics.

P. Szymański and T. Kajdanowicz. 2017. A scikit-based Python environment for performing multi-label classification. *ArXiv e-prints*.

Mingyang Wang, Heike Adel, Lukas Lange, Jannik Strötgen, and Hinrich Schütze. 2023. NLNDE at SemEval-2023 task 12: Adaptive pretraining and source language selection for low-resource multilingual sentiment analysis. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 488–497, Toronto, Canada. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing. *Preprint*, arXiv:1910.03771.