McGill-NLP at SemEval-2025 Task 11: Bridging the Gap in Text-Based Emotion Detection

Vivek Verma[1,2*], David Ifeoluwa Adelani[2,3,4]

[1]Université de Montréal, [2]Mila - Quebec AI Institute,
[3]McGill University [4]Canada CIFAR AI Chair
vivek.verma.1@umontreal.ca

Abstract

In this paper, we present the results of our SemEval-2025 Emotion Detection Shared Task Track A which focuses on multi-label emotion detection. Our team's approach leverages prompting GPT-4o, fine-tuning NLLB-LLM2Vec encoder, and an ensemble of these two approaches to solve Track A. Our ensemble method beats the baseline method that fine-tuned RemBERT encoder in 24 of the 28 languages. Furthermore, our results shows that the average performance is much worse for under-resourced languages in the Afro-Asiatic, Niger-Congo and Austronesia with performance scores at 50 F1 points and below. Our simple approach ranked second for Kinyarwanda, and ranked in top-5 for Afrikaans, Algerian Arabic, Nigerian-Pidgin, Sundanese, and Swahili, in Track A of the Emotion Detection Shared Task. [1]

1 Introduction

Emotion detection is the task of identifying and categorizing emotions expressed in textual data (Acheampong et al., 2020). Given a piece of text, such as sentence, document, sentence, or summary, the goal could be to determine the underlying emotion conveyed by the speaker or that invoked in the listener. Emotions play an essential role human interaction and interpersonal relationships (Ekman, 1999). In actual speech, speakers often give many non-verbal cues such as facial expressions or hand gestures to convey their emotions. Even then emotions of the speaker can be hard to detect and this problem becomes even harder for text because of the subtle cues, lack of emoticons, sarcasm or satire, or complexity and ambiguity of language (Kratzwald et al., 2018; Chatterjee et al., 2019).

People use text on social media such as reddit and there is a growth of textual dialogue with growing prominence of these platforms (Chatterjee et al., 2019). Most human-LLM interaction is in text and hence Emotion detection in text is important for NLP models to respond appropriately. Nandwani and Verma (2021) describe a wide variety of cases for businesses, healthcare sector, education sector, etc. that have a need for accurate emotion detection.

Wide array of approaches have been tried to solve emotion detection. Acheampong et al. (2020) provide a list of state of the art approaches from 2015 to 2020 including rule-based, machine learning based, and hybrid approaches that combine the two. In SemEval2018 (Mohammad et al., 2018), Alhuzali and Ananiadou (2021) use BERT (Devlin et al., 2019) to learn contextualized word representation and combine it with a linear layer to get a single score for each token to solve multi-label emotion detection in English, Arabic, and Spanish. For the same task Huang et al. (2021) use bi-directional LSTMs as encoders and decoders. Das et al. (2021) experiment with various machine learning, deep learning, and transformer based architectures and get best performance from XLM-R (Conneau et al., 2020) for emotion classification on Bengali text. Plaza-del Arco et al. (2022) explore emotion classification in zero shot learning setup with Natural Language Inference (NLI). In more recent times, LLMs are being used for all sorts of NLP tasks and at SemEval-2024 Task 3 (Wang et al., 2024), a third of the teams used LLMs. Team *petkatz* (Kazakov et al., 2024) that ranked second, fine-tuned GPT-3.5 for emotion classification.

Our team used a combination of GPT-4o (OpenAI, 2024) given the success of GPT-4 (OpenAI et al., 2024), and NLLB-LLM2Vec (Schmidt et al., 2024) which integrates Machine Translation (MT) encoders into LLM backbones. MT-LLMs preserve the multilingual representation alignment from MT encoder, allowing low resource languages to tap into the knowledge of English-centric LLMs. For GPT-4o we use few-shot prompting technique

[*] This work was carried out during internship at Mila.

[1] Our code will be made available here

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1783–1789
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

(Brown et al., 2020) and with NLLB-LLM2Vec we fine-tune LoRA adapters for each language with the training and dev set of the data provided.

2 Task Description

The task (Muhammad et al., 2025b) focuses on identifying and quantifying perceived emotion of the speaker based on a given sentence or a short text snippet. This means determining the emotion that the speaker appears to express, rather than the emotions invoked in the listerner, or anyone else mentioned in the text, or the true emotion of the speaker.

The emotions being looked at for this task are from Ekman's six basic emotions (Ekman, 1992) - joy, sadness, fear, anger, surprise, or disgust. The definition of these emotions is provided in Muhammad et al. (2025a), which we present here:

- Joy: "Expressions of happiness, pleasure, or contentment."
- Sadness: " Expressions of unhappiness, sorrow, or disappointment."
- Fear: "Expressions of anxiety, apprehension, or dread."
- Anger: "Expressions of frustration, irritation, or rage."
- Surprise: "Expressions of astonishment or unexpected events."
- Disgust: "A reaction to something offensive or unpleasant."

The shared tasks has **Three Tracks** of problems - Track A, Track B, and Track C. **Track A** is for multi-label emotion detection where given a piece of text we evaluate the presence of each of the six emotions described above and give a binary classification of 0 or 1 for each of them. **Track B** is for emotion intensity detection where each of the emotions can have ordinal values from 0 to 3. 0 meaning no emotion, 1 for low degree of emotion, 2 for moderate degree of emotion, and 3 for high degree of emotion. **Track C** is for Cross-lingual emotion detection where given labeled training data in a language, predictions need to be done for text instances in another language. Some languages in all the 3 tasks include five emotions, excluding the emotion *disgust*.

We focus on a single track, so our submission is only for Track A which had 28 languages in the dataset (Muhammad et al., 2025a; Belay et al., 2025): Afrikaans (afr), Algerian Arabic (arq), Amharic (amh), Portuguese (Brazilian) (ptbr), Mandarin Chinese (chn), Emakhuwa (vmw), English (eng), German (deu), Hausa (hau), Hindi (hin), Igbo (ibo), Kinyarwanda (kin), Spanish (Latin American) (esp), Marathi (mar), Moroccan Arabic (ary), Portuguese (Mozambican) (ptMZ), Nigerian-Pidgin (pcm), Oromo (orm), Romanian (ron), Russian (rus), Somali (som), Sundanese (sun), Swahili (swa), Swedish (swe), Tatar (tat), Tigrinya (tir), Ukrainian (ukr), Yoruba (yor).

3 System Overview

Our system for solving Track A consists of GPT-4o, NLLB-LLM2Vec and a method of ensemble to combine the best results from the two. We describe each of these below:

3.1 Prompting GPT-4o in few-shot setting

We prompt GPT-4o via API. For each language the prompt has a static context part that consists of Task description, the order of emotions, and first 50 examples from the training set. A dynamic query is added to the static context for each entry in the test set, requiring one API call for each entry.

The DEV set is used heuristically, before running on the test set, to try different prompts and prompt structures, and chose one that gives the best F1-score. For all languages, we kept the prompt in English and only use examples and test case in the target language. We stick to prompting in English since previous work already suggest that prompting in English works better on average (Lin et al., 2021).

3.2 NLLB - LLM2Vec

NLLB-LLM2Vec was developed to combine the excellent multilingual representations of MT models with LLMs that excel on English NLU, due to their language modeling training on large corpora. It fuses NLLB 600M parameter model MT encoder (Team et al., 2022) and Llama 3-8B variant (Meta AI, 2024) that underwent 'LLM2Vec process' (BehnamGhader et al., 2024)

We fine-tune LoRA adapters for each language. We train with rank r=16, alpha α=32, and use 4-bit QLoRA-style quantization (Dettmers et al., 2023). We use weight decay of 0.01, per device batch size

Language	**GPT-4o**	**NLLB-LLM2Vec**	**Ensemble**	**Baseline**
Afrikaans (afr)	60.1	32.5	**60.1**	37.1
Amharic (amh)	51.3	63.0	63.4	63.8
Algerian Arabic (arq)	57.0	41.0	**58.7**	41.4
Moroccan Arabic (ary)	51.4	39.8	**52.5**	47.2
Chinese (chn)	54.9	55.9	**59.6**	53.1
German (deu)	63.8	62.0	**64.4**	64.2
English (eng)	73.6	77.5	**77.7**	70.8
Spanish (esp)	79.1	72.1	**79.1**	77.4
Hausa (hau)	64.3	57.8	**65.4**	59.6
Hindi (hin)	84.3	87.0	**88.0**	85.5
Igbo (ibo)	52.2	48.7	**53.2**	47.9
Kinyarwanda (kin)	54.9	40.3	**58.9**	46.3
Marathi (mar)	87.1	85.4	**87.5**	82.2
Oromo (orm)	49.9	48.2	**56.4**	12.6
Nigerian-Pidgin (pcm)	55.9	63.2	**63.2**	55.5
Portuguese (pt-br)	56.8	40.3	**56.8**	42.6
Portuguese (pt-MZ)	40.6	42.2	45.1	45.9
Romanian (ron)	69.8	67.2	72.8	76.2
Russian (rus)	83.8	84.4	**86.4**	83.8
Somali (som),	48.4	40.7	**48.4**	45.9
Sundanese (sun)	50.3	26.6	**50.3**	37.3
Swahili (swa)	32.0	21.8	**35.9**	22.7
Swedish (swe)	50.8	42.4	51.1	52
Tatar (tat)	73.9	39.3	**73.9**	53.9
Tigrinya (tir)	42.5	44.3	**48.2**	46.3
Ukrainian (ukr)	60.0	37.5	**60.0**	53.5
Emakhuwa (vmw)	12.3	6.0	**12.6**	12.1
Yoruba (yor)	33.0	19.5	**33.0**	9.2
Average F1	**56.9**	**49.6**	**59.4**	**50.9**

Table 1: Result on test set for GPT-4o (few-shot), NLLB-LLM2Vec, and Ensemble of these two (Ranked Submission), along with the Baseline RemBERT score. Instances where our Ensemble method does better than Baseline RemBERT are marked in bold. The average macro F1 across all languages for each system is shown in the last row.

of 4, and train for 2 epochs. We use the provided train set and dev set entirely as training and validation sets. After fine-tuning we scan for a threshold between 0.3 to 0.7 in increments of 0.05 for classification so that it maximizes f1 score on validation set. We then generate predictions on test set from fine-tuned model. Fine-tuning was done on a single Nvidia V100 32GB GPU.

3.3 Ensemble of GPT-4o and NLLB-LLM2Vec

For our final submission, we use the best results from the two systems at an emotion level. This means that a submission for a single language could have a few emotions from GPT-4o and a few from NLLB-LLM2Vec based on which of the two performed better for each emotion. Ideally, this should be done looking at those results on the validation set and predecided, instead of looking at test set results and choosing the best one.

4 Results and Discussion

The results from the two approaches, and the ensemble which is our final ranked submission, are shown in the Table 1. The performance varies from 0.1256 macro F1 for Emakhuwa (vmw) being the lowest to 0.8802 for Hindi (hin) being the highest. We computed the results for all the 28 languages. Comparing our results to other teams, on 12 languages out of 28, our team is in the top 10 rankings for that language. Our ensemble method does better than the Baseline RemBERT (Chung et al., 2021) provided by the Task organizers (Muhammad et al., 2025a) on 24 languages out of 28.

Language Family	**Average ranked F1 score**
Afro-Asiatic	50.7
Austronesian	50.3
Creole	63.2
Indo-European	69.1
Niger-Congo	45.3
Sino-Tibetan	59.6
Turkic	73.9

Table 2: Average macro F1 score of the ranked submission, grouped by language family.

Average F1 score across language families (Muhammad et al., 2025a; Belay et al., 2025) is shown in the Table 2. We see that Turkic and Indo-European languages have the highest scores while Austronesian and Niger-Congo languages have the lowest scores. Few-shot prompting on GPT-4o performs better than NLLB-LLM2Vec on most languages in the Indo-European, Niger-Congo, Turkic, and Austronesian families, whereas NLLB-LLM2Vec performs better on languages in the Creole, and Sino-Tibetan family and half the languages in Afro-Asiatic family.

There are a few things we could've done to look for better performance. We have the same training parameters for NLLB-LLM2Vec for each language and we might have benefited from tuning parameters at a language level. We could've also increased the number of epochs of training for better performance.

4.1 Experiments Post Competition Deadline

Given our shortcomings in the fine-tuning of our NLLB-LLM2Vec, we evaluate the languages more granularly to improve performance. We introduce lora dropout of 0.05 in the LoRA parameters, we train for 5 epochs instead of 2 and chose the epoch that maximizes the F1 score on the dev set. At each epoch, we generate classification probabilities for each emotion in the dev set. To convert the probabilities to classification, we need a threshold

Language	**NLLB-LLM2Vec**	**NLLB-LLM2Vec (Tuned)**
Afrikaans (afr)	32.5	**47.1**
Amharic (amh)	63.0	66.0
Algerian Arabic (arq)	41.0	45.8
Moroccan Arabic (ary)	39.8	**47.7**
Mandarin Chinese (chn)	55.9	**62.5**
German (deu)	62.0	62.9
English (eng)	77.5	77.7
Spanish (esp)	72.1	76.1
Hausa (hau)	57.8	59.5
Hindi (hin)	87.0	87.3
Igbo (ibo)	48.7	49.0
Kinyarwanda (kin)	40.3	**47.8**
Marathi (mar)	85.4	81.9
Oromo (orm)	48.2	**55.0**
Nigerian-Pidgin (pcm)	63.2	64.1
Portuguese (Brazilian) (ptbr)	40.3	**48.8**
Portuguese (Mozambican) (ptmz)	42.2	45.2
Romanian (ron)	67.2	68.9
Russian (rus)	84.4	84.7
Somali (som)	40.7	43.2
Sundanese (sun)	26.6	**42.0**
Swahili (swa)	21.8	27.6
Swedish (swe)	42.4	**49.7**
Tatar (tat)	39.3	**60.8**
Tigrinya (tir)	44.3	**53.0**
Ukrainian (ukr)	37.5	**44.7**
Emakhuwa (vmw)	6.0	**28.4**
Yoruba (yor)	19.5	**34.5**
Average F1	**49.6**	**55.8**

Table 3: Results on test set for NLLB-LLM2Vec (Tuned) with better hyperparameter tuning (Post Competition Deadline) compared to that of the results in the ranked submission before competition deadline. We see significant performance gain on the test set compared to our previous iteration for NLLB-LLM2Vec, with macro F1 average increasing from 49.6 to 55.8. This improvement surpasses the Baseline macro F1 average of 50.9. Languages with significant improvement due to hyperparameter tuning are marked in bold.

cut-off, and we choose to have a different threshold for each emotion. We do this by sweeping through threshold values in the range [0.05, 0.7], in increments of 0.05, and selecting the one that maximizes the macro F1 score on the dev set. We use the test set prediction from the epoch with the best F1 score on dev set.

Our dedicated effort for languages improves performance for NLLB-LLM2Vec. We only fine-tune on dev set without looking at test set and score the test set a single time at the end. The new scores are shown in Table 3

We see that the average macro F1 score increased from 49.6 to 55.8 which is slightly behind GPT-4o with an average macro F1 score of 56.9. NLLB-LLM2Vec alone, without any ensemble technique, does better than Baseline RemBERT score on 19 languages out of 28 and on macro F1 average. On a few languages it does better than the previous Ensemble method, with the most noteworthy jump being on Emakhuwa (vmw) from 12.6 to 28.4. This would've placed our team on the second spot for this language. In about half of the languages there is a jump of greater than 5 points in F1 score, highlighting the importance of better hyperparameter tuning.

5 Conclusion

In this work, we present our approach for Track A of SemEval-2025 Task 11 for multi-label emotion detection in text on 28 languages. We experiment with two methods - few-shot prompting with GPT-4o and fine-tuning NLLB-LLM2Vec with LoRA adapters, and present an ensemble of these two for our best results. Our model outperforms Baseline results in most languages. We also see that average F1 scores are at 50 F1 points and below for under-resourced languages in the Afro-Asiatic, Niger-Congo and Austronesia.

We also present improvements and analysis on the performance of NLLB-LLM2Vec for this task with better hyperparameter tuning on the dev set. This leads to NLLB-LLM2Vec beating Baseline results in 19 languages, and its performance gets close to GPT-4o performance on average macro F1 score across languages. Further work can be done to improve the Ensemble technique to work more fluently and combine the results of GPT-4o and NLLB-LLM2Vec based on the dev set. This would invariably provide better results than our ranked submission.

References

Frank A. Acheampong, Wenyu Chen, and Henry Nunoo-Mensah. 2020. Text-based emotion detection: Advances, challenges, and opportunities. *Engineering Reports*, 2(e12189).

Hassan Alhuzali and Sophia Ananiadou. 2021. SpanEmo: Casting multi-label emotion classification as span-prediction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1573–1584, Online. Association for Computational Linguistics.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. Llm2vec: Large language models are secretly powerful text encoders. *Preprint*, arXiv:2404.05961.

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.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Ankush Chatterjee, Umang Gupta, Manoj Kumar Chinnakotla, Radhakrishnan Srikanth, Michel Galley, and Puneet Agrawal. 2019. Understanding emotions in text using deep learning and big data. *Computers in Human Behavior*, 93:309–317.

Hyung Won Chung, Thibault Fevry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2021. Rethinking embedding coupling in pre-trained language models. In *International Conference on Learning Representations*.

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

Avishek Das, Omar Sharif, Mohammed Moshiul Hoque, and Iqbal H. Sarker. 2021. Emotion classification in a resource constrained language using transformer-based approach. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 150–158, Online. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Thirty-seventh Conference on Neural Information Processing Systems*.

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.

Paul Ekman. 1992. Are there basic emotions? *Psychological Review*, 99(3):550–553.

Paul Ekman. 1999. Basic emotions. In Tim Dalgleish and Mick J. Power, editors, *Handbook of Cognition and Emotion*, pages 45–60. John Wiley & Sons Ltd.

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.

Roman Kazakov, Kseniia Petukhova, and Ekaterina Kochmar. 2024. PetKaz at SemEval-2024 task 3: Advancing emotion classification with an LLM for emotion-cause pair extraction in conversations. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1127–1134, Mexico City, Mexico. Association for Computational Linguistics.

Bernhard Kratzwald, Suzana Ilić, Mathias Kraus, Stefan Feuerriegel, and Helmut Prendinger. 2018. Deep learning for affective computing: Text-based emotion recognition in decision support. *Decision Support Systems*, 115:24–35.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona T. Diab, Veselin Stoyanov, and Xian Li. 2021. Few-shot learning with multilingual language models. *CoRR*, abs/2112.10668.

Meta AI. 2024. Llama 3 model card.

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.

Pankaj Nandwani and Raksha Verma. 2021. A review on sentiment analysis and emotion detection from text. *Social Network Analysis and Mining*, 11(1):81.

OpenAI. 2024. Hello gpt-4o. Online. Accessed: 2025-02-25.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Flor Miriam Plaza-del Arco, María-Teresa Martín-Valdivia, and Roman Klinger. 2022. Natural language inference prompts for zero-shot emotion classification in text across corpora. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6805–6817, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Fabian David Schmidt, Philipp Borchert, Ivan Vulić, and Goran Glavaš. 2024. Self-distillation for model stacking unlocks cross-lingual nlu in 200+ languages. *Preprint*, arXiv:2406.12739.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Fanfan Wang, Heqing Ma, Rui Xia, Jianfei Yu, and Erik Cambria. 2024. SemEval-2024 task 3: Multimodal emotion cause analysis in conversations. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2039–2050, Mexico City, Mexico. Association for Computational Linguistics.