Deepwave at SemEval-2025 Task 11: Emotion Analysis in Low-Resource Settings Using LLM and Data Augmentation

Shenpo Dong[1], **Zhilong Ji**[1],
[1]Tomorrow Advancing Life
Correspondence: dongshenpo@tal.com , jizhilong@tal.com

Abstract

This paper introduces a new emotion detection method designed for low-resource languages, specifically for the SemEval-2025 Task 11 challenge. The approach fine-tunes Google's Gemma 2 model using Chain-of-Thought prompting augmentation data. The methodology integrates supervised fine-tuning and model ensembling, leading to substantial improvements in multi-label emotion recognition, emotion intensity prediction, and cross-lingual performance. The results demonstrate robust performance across various low-resource language scenarios. On task A, our method achieves an average improvement of 6.96 F1. On task B, it yields an average increase of 23.3 F1. For task c, the proposed approach improves metrics for low-resource language families by 50% to 70%.

1 Introduction

Text sentiment analysis, a cornerstone of natural language processing (NLP), aims to computationally identify and extract subjective information from text, such as opinions, emotions, and attitudes. Over the years, this field has evolved significantly, driven by the need to understand human sentiment in various domains, including customer feedback, social networks, and product reviews (Liu and Chen, 2015). Traditional methods, such as lexicon-based approaches and machine learning models, have laid the foundation for sentiment analysis (Wiebe et al., 2005; Salam and Gupta, 2018). However, the advent of large language models (LLMs) has revolutionized this field, offering new capabilities and challenges.

Early sentiment analysis methods relied heavily on lexicon-based techniques, where predefined sentiment scores were assigned to words, and heuristic rules were applied to aggregate these scores for an overall sentiment judgment. While these methods are computationally efficient, they often struggle with complex linguistic phenomena such as sarcasm, negation, and context-dependent sentiment. Machine learning models, like Naive Bayes, Support Vector Machines (SVM) , introduced a data-driven approach to sentiment analysis (Liu et al., 2017; Islam et al., 2022). These models were trained on labeled datasets to classify text into positive, negative, or neutral categories. However, their performance was limited by the need for extensive labeled data and their inability to capture deep semantic relationships within the text.

Large Language Models (LLMs) like BERT (Devlin et al., 2019) and ChatGPT (Ouyang et al., 2022) have revolutionized sentiment analysis, particularly in high-resource languages such as English. This advancement is largely attributed to their ability to achieve superior contextual understanding and facilitate zero-shot and few-shot learning through fine-tuning (Ameer et al., 2023) and prompt-based methods (Yu et al., 2022). Specifically: (1) Zero-Shot and Few-Shot Learning: LLMs' pre-trained language understanding enables them to perform sentiment analysis with minimal or no labeled data (Kuila and Sarkar, 2024). (2) Cross-Lingual Transfer: Multilingual pre-training (Yang et al., 2025) allows LLMs to transfer knowledge from high-resource to low-resource languages. (3) Prompt Engineering: Well-designed prompts guide LLMs to better interpret emotional expressions, particularly in low-resource languages.

However, despite their success in high-resource scenarios, LLMs face significant challenges when applied to low-resource languages (Barnes, 2023), particularly those in Africa and Southeast Asia. These challenges include: (1) Data Scarcity: The limited availability of labeled data in low-resource languages hinders the training of traditional supervised learning methods (Belay et al., 2025). (2) Linguistic Diversity: The complex syntax and diverse dialects present in these languages complicate model understanding and generalization. (3) Cul-

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 1957–1963

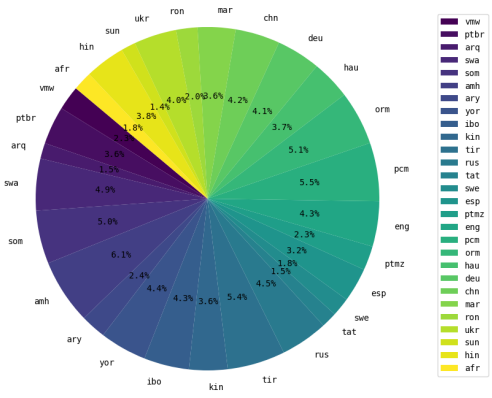

Figure 1: BRIGHTER languages distribution

tural Differences: The variations in emotional expression across cultures necessitate cross-cultural understanding, which is often challenging for models trained primarily on high-resource language data (Tafreshi et al., 2024).

In analogous event extraction tasks, the classical pipeline approach is divided into two stages, event argument extraction and event relation extraction, to enhance end-to-end accuracy (Dong et al., 2022). Consequently, for low-resource scenarios, we draw upon this concept and propose a two-stage augmented data-based sentiment extraction method.

The main contributions are summarized as: 1. We introduce a two-stage CoT-enhanced data pipeline, which generates interpretable and diverse English augmented data for low-resource languages to assist in model training. 2. We employ techniques such as supervised fine-tuning (SFT), K-fold cross-validation, model ensembling, and specialized LoRA adaptations. Across all three tasks, our approach achieves substantially higher performance compared to the baselines.

2 BRIGHTER Dataset

Understanding how emotions are expressed differently across languages is crucial for building inclusive digital tools. Muhammad et al. (2025a); Belay et al. (2025) have developed BRIGHTER, a comprehensive dataset encompassing 28 languages. BRIGHTER consists of 28 distinct datasets , each tailored to a specific language, designed to capture the nuanced expressions of emotions in text. These datasets are derived from a variety of sources, including social media posts, speeches, literary works, and news articles, ensuring a diverse representation of language usage. For some languages, new datasets were created, while existing ones were enhanced with automatically translated or generated data.

Each text instance within BRIGHTER is multi-labeled, indicating the presence of one or more of six core emotions, along with a neutral category. Furthermore, each emotion label is accompanied by an intensity rating on a 4-point scale, providing a more granular understanding of emotional expression. Analysis of BRIGHTER revealed that emotion recognition remains a significant challenge for Large Language Models (LLMs), particularly for languages with limited resources.

SemEval Task 11 contains three tasks for emotion analysis (Muhammad et al., 2025b): Task A focuses on multi-label emotion detection, classifying text snippets into six emotions (joy, sadness, fear, anger, surprise, disgust), with varying presence of the "disgust" label across languages. Task B involves emotion intensity prediction, assigning ordinal intensity levels (0-3) to given text and emotion pairs. Task C tackles cross-lingual emotion detection, requiring prediction of emotion labels in a target language using training data from a different language.

3 System Overview

In this study, we propose a novel approach for text sentiment analysis using large language models through Supervised Fine-Tuning (SFT), Chain-of-Thought (CoT) prompting (Wei et al., 2022), data augmentation and model ensemble techniques. Our methodology integrates advanced LLM capabilities with traditional sentiment analysis frameworks to achieve higher accuracy and robustness in sentiment classification tasks.

3.1 Chain-of-Thought

To enhance the model's capacity for sentiment analysis, particularly in cross-lingual settings, we employed Chain-of-Thought (CoT) prompting to generate explicit intermediate reasoning steps. This approach facilitates a more nuanced understanding of sentiment, especially in contexts where it is implicitly conveyed.

Initial pilot experiments were conducted on high-resource languages, specifically English and Chinese, to validate the efficacy of CoT in this domain. The Gemma 2 27B IT (Team, 2024) model was utilized to implement the CoT framework, decomposing the sentiment analysis task into two distinct sub-tasks:

Speaker	I can't move, my hand is stuck, I'm making weird noises and my mom is screaming.
Keyword Identification	"can't move": helplessness, fear, panic, entrapment; "stuck": helplessness, frustration, panic, entrapment; "weird noises": fear (of the unknown, of one's own body), confusion, distress; "mom is screaming": fear, alarm, panic, distress
Sentiment Recognition	Let's classify the emotions based on the given categories, allowing for multiple options: "can't move": Fear: A very strong primary emotion here, due to the physical inability and potential danger. Surprise: The sudden inability to move could certainly evoke surprise...

(i) sentiment keyword identification. *Prompt: "What emotional keywords are included in this sentence and output them in JSON format."* Like entity recognition, the LLM first analyzes sentence to identify emotionally expressive words. Without predefined category knowledge, the LLM generates varied emotional expressions, like alarm, panic, resulting in what we term coarse-grained data.

(ii) sentiment polarity recognition. *Prompt: "Summarize these emotions, with candidates including 'anger', 'fear', 'joy', 'sadness',' surprise', and' disgust'. You can choose from multiple options."* In this task, we utilize the LLM to map the previously extracted emotional terms to five predefined labels. The emotional labels in this data now perfectly align with BRIGHTER's label taxonomy, where each emotion category has one and only one standardized description. We therefore classify this as fine-grained data.

This augmented data was designed to provide the model with a granular understanding of the intricate relationship between specific keywords and their associated sentiments, thereby improving the model's overall sentiment analysis performance. We performed the same procedure on low-resource languages to generate augmented data. Notably, in our augmented dataset, all responses were strictly required to be in English except for the keywords.

We keep data matching ground truth after Task i and ii. For mismatches, we raise sample temperature and re-run the tasks iteratively until 80% of data has both coarse- and fine-grained data. This iterative process aimed to enhance the quality and diversity of the augmented dataset. For subsequent tasks, we consistently train the models using both the augmented data and the original data in combination.

3.2 Fine-Tuning

To maintain computational efficiency during the supervised fine-tuning (SFT) phase, we selected the Gemma 2 9B IT model as our base model. This decision was driven by the model's balance of performance and resource requirements, enabling us to conduct extensive experimentation within feasible time constraints.

Addressing the challenge of limited data availability, which is particularly prevalent in Tasks 1 and 2 and often leads to suboptimal model performance, we implemented a K-fold cross-validation training strategy. Specifically, we set K to 5, dividing our dataset into five equal partitions. This dataset comprised a mixture of coarse-grained sentiment data from Task 1 and fine-grained sentiment data from Task 2, effectively creating a unified training corpus. We then iteratively trained five distinct models, with each model trained on four partitions and validated on the remaining partition. This cross-validation approach allowed us to maximize the utilization of our limited data while simultaneously providing a robust estimate of model performance and mitigating the risk of overfitting.

During the training process within each fold of the cross-validation, we employed the macro-averaged F1-score as the primary evaluation metric for each language. This metric provided a comprehensive assessment of the model's performance across all classes within a given language.

$$F1_{macro}(l) = \frac{1}{|C|} \sum_{c \in C} \frac{2 \cdot p_c \cdot r_c}{p_c + r_c}$$

where: $|C|$ represents the number of classes in the language. p_c and r_c are the precision and recall for class c, respectively. To select the optimal checkpoint for each fold, we calculated the average of the macro-averaged F1-scores across all languages. This average score served as the criterion for checkpoint selection.

$$Score = \frac{1}{|L|} \sum_{l \in L} F1_{macro}(l)$$

Language	afr	arq	ary	chn	deu	eng	esp	hau	hin	ibo	kin	mar	pcm	ptbr	ptmz	rus	sun	swa	swe	tat	ukr	vmw	yor	AVG
XLM-R*	10.82	31.98	40.66	58.48	55.37	67.3	29.85	36.95	33.71	18.36	32.93	78.95	52.03	15.4	30.72	78.76	19.66	22.71	34.63	26.48	17.77	9.92	11.94	35.45
mBERT*	25.87	41.75	36.87	49.61	46.78	58.26	54.41	47.33	54.11	37.23	35.61	60.01	48.42	32.05	14.81	61.81	27.88	22.99	44.24	43.49	31.74	10.28	21.03	39.41
Qwen2.5-72B*	**60.18**	37.78	**52.76**	55.23	59.17	55.72	72.33	43.79	79.73	37.4	31.96	74.58	38.66	51.6	40.44	73.08	42.67	27.36	48.89	51.58	54.76	**20.41**	24.99	49.35
Mixtral-8x7B*	53.69	45.29	35.07	44.91	51.2	58.12	65.72	40.4	62.19	31.9	26.35	50.36	45.61	41.64	36.52	61.72	42.1	26.51	48.61	39.44	40.15	19	19.67	42.87
DeepSeek-R1-70B*	43.66	50.87	47.21	53.45	54.26	56.99	73.29	51.91	76.91	32.85	32.52	76.68	45	51.49	39.58	76.97	**44.61**	**33.27**	44.6	53.86	51.19	19.09	**27.44**	49.46
Ours† (5-models-merge)	51.46	53.37	51.93	61.46	61.97	72.41	78.47	54.96	88.73	41.75	34.46	83.11	54.84	51.75	41.2	86.43	29.49	17.59	49.13	64.17	55.46	9.34	13.44	51.47
Ours (5-models-merge)	52.34	**58.2**	52.45	**62.01**	**65.55**	**75.11**	**79.43**	**59.4**	**90.13**	**48.07**	**37.97**	**87.81**	**59.45**	**53.86**	**45.09**	**87.25**	37.94	22.72	**56.9**	**68.21**	**61.89**	12.14	23.96	**56.42**

Table 1: Average F1-Macro for Task A Multi-label Emotion Recognition. The data marked with * represents that from (Muhammad et al., 2025a). The symbol † is used to denote models trained exclusively on the original dataset.

Lang	arq	chn	deu	eng	esp	ptbr	rus	ukr	AVG
XLM-R*	0	36.92	38.3	37.36	55.72	18.24	68.96	36.16	36.45
mBERT*	0	21.96	17.35	25.74	27.94	8.36	37.63	4.32	17.91
Qwen2.5-72B*	29.54	46.17	43.3	55.99	51.11	38.2	58.25	37.74	45.03
Mixtral-8x7B*	31.05	46.52	47.6	55.26	55.54	39.17	56.01	38.74	46.23
DeepSeek-R1-70B*	36.37	48.57	54.78	48.08	60.74	46.72	62.28	43.54	50.13
Ours†(5-models-merge)	51.22	**71.96**	66.91	79.35	74.94	61.37	87.74	54.17	67.46
Ours (5-models-merge)	**57.41**	69.42	**74.24**	**80.91**	**79.21**	**68.41**	**91.64**	**66.24**	**73.43**

Table 2: Average F1-Macro for Task B Emotion Intensity. The data marked with * represents that from (Muhammad et al., 2025a). The symbol † is used to denote models trained exclusively on the original dataset.

Lang	afr	arq	ary	chn	deu	eng	esp	hau	hin	ibo	mar	pcnr	ptbr	ptm	ron	rus	sun	swe	tat	ukr	vmv	yor	zul	AVG
mBERT*	16.95	31.38	24.83	21.61	28.6	18.8	30.09	15.59	36.94	9.94	42.32	22.55	23.86	13.54	61.5	37.15	25.29	28.86	35.81	25.69	12.11	9.62	13.04	20.93
mDeBERTa*	33.25	35.92	36.28	42.41	42.61	35.3	37.09	32.8	57.74	9.52	54.05	25.39	34.42	24.46	60.6	29.7	27.31	43.28	47.72	35.12	11.74	10.03	13.87	27.87
LaBSE*	35.12	35.93	42.83	45.28	42.45	36.71	54.56	38.46	69.78	18.13	74.65	33.29	41.51	31.44	69.79	61.32	34.79	44.24	60.66	44.37	9.65	11.64	18.16	34.09
Ours † (lora)	41.33	35.46	49.19	64.74	65.72	78.97	76.49	63.42	89.77	**57.49**	81.62	61.91	54.76	51.33	71.46	81.76	**48.15**	56.74	71.47	62.01	12.10	27.46	7.16	43.54
Ours (lora)	**57.41**	**58.75**	**63.22**	**68.89**	**72.67**	**79.69**	**83.11**	**70.88**	**91.87**	55.35	**90.29**	**67.4**	**62.91**	**55.54**	**76.7**	**90.58**	46.66	**64.53**	**78.86**	**70.18**	**21.04**	**34.16**	**19.27**	**52.85**

Table 3: Average F1-Macro for Task C Crosslingual Multi-Label Classification. The data marked with * represents that from (Muhammad et al., 2025a). The symbol † is used to denote models trained exclusively on the original dataset.

	Romance	Germanic	Semitic	Niger-Congo	Slavic	Sino-Tibetan
mBERT*	32.25	23.30	23.93	16.33	32.88	21.61
mDeBERTa*	39.14	38.61	35.00	20.58	37.51	42.41
LaBSE*	49.33	39.63	39.07	25.47	55.45	45.28
Ours † (lora)	63.51	60.69	49.36	38.80	71.75	64.74
Ours (lora)	**69.57**	**68.58**	**64.28**	**44.34**	**79.87**	**68.89**

Table 4: Average F1-Macro across Language Families in Task C.

3.3 Model Ensemble

To further enhance the performance of our sentiment analysis model, we employed a model ensembling technique. Specifically, we utilized LLM merging (Goddard et al., 2024), a strategy that combines the predictions of multiple LLM to achieve improved generalization. Given that the five models generated through our K-fold cross-validation were derived from the same base architecture and exhibited comparable performance on their respective validation sets, we opted for a linear weighted merging approach.

We perform a linear fusion of the five models into a single unified model to achieve an optimal tradeoff between inference speed and model performance. As such, each model was assigned a merge weight of 0.2, ensuring an equal contribution to the final ensemble prediction. This straightforward yet effective technique was chosen for its ability to reduce variance and mitigate the risk of overfitting, ultimately leading to a more robust and reliable sentiment analysis system.

3.4 Crosslingual Recognition

To achieve the objective of cross-lingual detection in Task C, we employed a straightforward yet effective strategy involving the training of multiple Low-Rank Adaptation (LoRA) modules. Specifically, for each target language l_i, a dedicated LoRA (Hu et al., 2022) module was trained. The training dataset for this module comprised data from all other languages within Task A, excluding the target language l_i. This approach facilitated the model's ability to discern subtle linguistic nuances and patterns characteristic of languages distinct from l_i.

Given the inherent time-intensive nature of training individual LoRA modules for each language, we abstained from employing model fusion techniques in Task C.

4 Experimental Setup

To maximize coverage for multilingual tasks, we selected the Gemma 2 multilingual model (Team, 2024) as the base model for augmenting and training. We employed bfloat16 precision and Adam optimizer was configured with beta values of 0.9 and 0.999, and an epsilon of 1e-8. To optimize the training process, we set the learning rate to 7e-6. The batch size was configured to 64, which allowed for efficient utilization of computational resources while maintaining reasonable training speed. To ensure robust evaluation and mitigate overfitting, we adopted a 5-fold cross-validation strategy. For both Task A and Task B, we trained the models using the complete set of original data along with the augmented data, and employed Mergekit (Goddard et al., 2024) for linear model fusion. For Task C, we continue to employ mixed data while training dedicated LoRA heads for each individual language.

5 Results

Our proposed work demonstrates significant improvements across three tasks, leveraging Gemma's cross-lingual capabilities enhanced through chain-of-thought reasoning, data augmentation, and model-ensembling strategies.

In task A Multi-label Emotion Recognition (Table 1), Deep achieves superior performance with an average F1-macro score of 56.42, outperforming all comparable systems by +6.96 points over the strongest baseline (Deepseek R1-70B: 49.46). Notably, in certain low-resource language scenarios, the performance metrics substantially surpass those of the English context: +13.21 in Hindi (90.13 vs. 76.91), +11.13 in Marathi (87.81 vs. 76.68) and +10.28 in Russian (87.25 vs. 76.97). However, there are also some scenes that perform poorly like Emakhuwa.

For Task B Emotion Intensity, which requires simultaneous prediction of both emotion categories and intensity levels, our method demonstrates superior performance in Table 2. Our method attaining 73.43 average F1-score +23.3 points higher than DeepSeek-R1-70B (50.13). Our method demonstrates superior performance over the baseline across nearly all language scenarios, as exemplified by the following cases: achieving 91.64 in Russian and 79.21.

For task C, Crosslingual Multi-Label Classification (Table 3), we establishe a new state-of-the-art performance with 52.85 average F1-score, surpassing previous best results by +18.76 points (LaBSE*: 34.09). As observed in Table 4, our method achieves nearly 60-70% improvements for low-resource language families (e.g., Niger-Congo and Semitic), demonstrating remarkable effectiveness in data-scarce scenarios.

As shown in Tables Table 1- 4, we conducted an additional experiment using identical training configurations to our final approach, with the sole variation being the training data. Model marked with † were trained exclusively on the original data, while final model utilized both original and augmented data. The results demonstrate that the hybrid data approach (original + augmented) consistently outperforms the original-data-only by an average margin of 6-10 percentage points. This also indicates that our augmented data underwent rigorous quality filtering, and its integration during training did not excessively interfere with the original data. On the contrary, by more explicitly highlighting the relationships between sentiment keywords and sentiment categories/intensities, it contributed to improved final performance.

6 Limitations

Our cross-lingual sentiment detection approach showed strong performance in Task 3, but its effectiveness was notably reduced in Tasks 1 and 2. This disparity is likely due to the training process involving datasets with both coarse and fine-grained sentiment annotations. The inherent differences in annotation granularity introduced inconsistencies, hindering the model's ability to accurately capture the subtle nuances of sentiment in these tasks.

Furthermore, the model performance in certain low-resource scenarios, particularly with Javanese, fell significantly below acceptable levels. This underscores the persistent challenge of adapting large language models to languages with limited available data. Future research should prioritize strategies for data augmentation in low-resource settings, such as back-translation and synthetic data generation, as well as the integration of language-specific linguistic resources. Additionally, exploring methods for efficient knowledge transfer and adaptation from high-resource to low-resource languages is crucial for bridging the performance gap.

7 Conclusion

This study investigated text-based emotion detection in low-resource languages, utilizing the Google Gemma 2 large language model. The research employed data augmentation, Chain-of-Thought (CoT) prompting, and model ensembling techniques. The proposed approach achieved substantial performance gains across multilingual emotion detection and emotion intensity prediction task, outperforming state-of-the-art baselines. Further research is needed to explore more effective knowledge transfer methods from high-resource to low-resource languages.

Overall, this work highlights the potential of large language models to bridge the gap in text-based emotion detection, particularly through data augmentation for resource-scarce language families.

References

Iqra Ameer, Necva Bölücü, Muhammad Hammad Fahim Siddiqui, Burcu Can, Grigori Sidorov, and Alexander Gelbukh. 2023. Multi-label emotion classification in texts using transfer learning. *Expert Systems with Applications*, 213:118534.

Jeremy Barnes. 2023. Sentiment and emotion classification in low-resource settings. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 290–304.

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.

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.

Shenpo Dong, Wei Yu, Hongkui Tu, Xiaodong Wang, Yunyan Zhou, Haili Li, Jie Zhou, and Tao Chang. 2022. Argumentprompt: activating multi-category of information for event argument extraction with automatically generated prompts. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 311–323. Springer.

Charles Goddard, Shamane Siriwardhana, Malikeh Ehghaghi, Luke Meyers, Vladimir Karpukhin, Brian Benedict, Mark McQuade, and Jacob Solawetz. 2024. Arcee's MergeKit: A toolkit for merging large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 477–485, Miami, Florida, US. Association for Computational Linguistics.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Simon Islam, Animesh Chandra Roy, Mohammad Shamsul Arefin, and Sonia Afroz. 2022. Multi-label emotion classification of tweets using machine learning. In *Proceedings of the International Conference on Big Data, IoT, and Machine Learning: BIM 2021*, pages 705–722. Springer.

Alapan Kuila and Sudeshna Sarkar. 2024. Deciphering political entity sentiment in news with large language models: Zero-shot and few-shot strategies. *arXiv preprint arXiv:2404.04361*.

Shuhua Monica Liu and Jiun-Hung Chen. 2015. A multi-label classification based approach for sentiment classification. *Expert Systems with Applications*, 42(3):1083–1093.

Yang Liu, Jian-Wu Bi, and Zhi-Ping Fan. 2017. A method for multi-class sentiment classification based on an improved one-vs-one (ovo) strategy and the support vector machine (svm) algorithm. *Information Sciences*, 394:38–52.

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.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Shaikh Abdul Salam and Rajkumar Gupta. 2018. Emotion detection and recognition from text using machine learning. *Int. J. Comput. Sci. Eng.*, 6(6):341–345.

Shabnam Tafreshi, Shubham Vatsal, and Mona Diab. 2024. Emotion classification in low and moderate resource languages. *arXiv preprint arXiv:2402.18424*.

Gemma Team. 2024. Gemma.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39:165–210.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yang Yu, Dong Zhang, and Shoushan Li. 2022. Unified multi-modal pre-training for few-shot sentiment analysis with prompt-based learning. In *Proceedings of the 30th ACM international conference on multimedia*, pages 189–198.