Language Resources and Evaluation
https://doi.org/10.1007/s10579-025-09846-z

ORIGINAL PAPER

Examining emotions in English and translated Chinese children's literature: a bilingual emotion detection model based on LLMs

Yanjin Liu[1] · Sophia Yat Mei Lee[2] · Dechao Li[2]

Accepted: 13 May 2025

Abstract

This study investigates the Chinese-English bilingual emotion detection within the context of children's literature. The study utilizes a parallel corpus of classical Chinese-English children's literature and compiles a bilingual dataset of emotionally-labelled text. The dataset is then leveraged to fine-tune and evaluate the performance of various Large Language Models (LLMs). The results indicate that the GPT-4o model outperforms alternative LLMs, achieving an F1 Micro score of 0.779 and an F1 Macro score of 0.764 on the evaluation task. These findings substantiate the viability of cross-lingual emotion detection within this domain and underscore the importance of selecting appropriate pre-training techniques. Furthermore, this study addresses specific cross-cultural challenges inherent in bilingual emotion detection, elucidating the complexities posed by language-specific and culturally bound emotional expressions. This study contributes to the expanding body of literature on emotion recognition in multilingual contexts, particularly in relation to the analysis of affective content in cross-cultural translated children's literature, and provides insights for future investigations in this field.

Keywords Emotion detection · Bilingual emotion analysis · Large language models · Fine-tuning · Children's literature

✉ Dechao Li
dechao.li@polyu.edu.hk

Yanjin Liu
yanjin.liu@connect.polyu.hk

Sophia Yat Mei Lee
ym.lee@polyu.edu.hk

[1] Faculty of Humanities and Social Sciences, City University of Macau, Macau, China

[2] Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong, China

Published online: 20 June 2025

1 Introduction

The study of emotion analysis, referred to as the exploration of emotions within language sciences (Majid, 2012), has been ongoing since the 1880s (Johnson-Laird & Oatley, 1989; Love, 2007; Wilce, 2009). It aims to determine the emotional tone or feelings conveyed by written content, such as text from social media posts (Brynielsson et al., 2014; Kohout et al., 2023), reviews (Tang et al., 2009), news articles (Edwards, 1999; Lin et al., 2008), or any other forms of written communication (Rimé, 2009). In recent decades, natural language processing (NLP) has significantly advanced emotion analysis, particularly through the development of language-based emotion classification datasets (Demszky et al., 2020; Sosea & Caragea, 2020). These datasets, which encompass diverse sources, including tweets (Ghosh et al., 2023a, 2023b; Mohammad & Bravo-Marquez, 2017), news articles (Staiano & Guerini, 2014), and literary texts (Haider et al., 2020).

Among all these language-based emotion classification datasets, the domain of children's literature is currently gaining prominence as a significant focal point for emotion analysis research (Adukia et al., 2022; Kaya et al., 2017; Moruzi et al., 2017; Oberländer & Klinger, 2018). This is perhaps due to children's literature's unique capacity to encapsulate a wide spectrum of emotions (Nikolajeva, 2014), intricately woven into narratives designed to engage and resonate with young readers. Children's literature often serves as a rich repository of emotional expression, offering "detailed information both about the characters' physical appearance and about their emotions and thoughts" (Nodelman, 2008, p. 13). As a result, analysing emotions within children's literature offers a valuable opportunity to delve into the emotional landscapes that shape young minds and contribute to their cognitive and emotional development (Wang et al., 2015). Through emotion analysis, we can not only enhance our understanding of literary narratives but also enrich our insights into the emotional dynamics that influence children's perceptions, experiences, and growth.

However, as Schwieter and Ferreira (2017) assert, there have been few attempts to study emotion in translated children's literature to date. Translated texts pose unique challenges as they not only inherit the intricate cultural and linguistic complexities of the original work but also represent the norms and expectations of the target culture (Toury, 1995). Investigating emotions in bilingual contexts assumes significance, as it allows for in-depth linguistic exploration of the intricate cultural and linguistic intricacies in emotions that arise from language differences. Through the provision of a bilingual dataset, researchers can delve into the cross-linguistic and cross-cultural variations in the expression of emotions, thereby facilitating a deeper comprehension of emotion recognition and sentiment analysis within bilingual contexts. Moreover, the establishment of a unified system for detecting emotion categories across both English and Chinese languages holds immense value, serving as a valuable tool not only for future studies detecting emotions in Chinese-English language pairs but also for examining the nuances of translations and conducting comprehensive cross-cultural analyses.

To address the research gap, this paper attempts to compile a bilingual Chinese-English dataset of emotions. The dataset is composed of 2,177 Chinese and English sentences annotated with emotions taken out of a parallel Chinese-English children's literature dataset. The emotion taxonomy in this paper, adapted from Parrott (2001) five basic emotions (JOY, SADNESS, ANGER, FEAR, and LOVE), is designed considering the psychological and practical implication in the dataset. Subsequent steps involve deploying supervised machine learning techniques (Logistic Regression, Random Forest, Support Vector Machines), unsupervised deep learning with Neural Networks, and advanced Transformer models (XLM-RoBERTa, GPT family, DeepSeek-R1, and Qwen-2.5) for fine-tuning on labelled datasets. The fine-tuning results indicate that the GPT-4o model from GPT family consistently surpasses other models, reaching its best performance with an F1 Micro score of 0.779 and an F1 Macro score of 0.764.

The contributions of this paper include: (1) it introduces BilingualChildEmo, a novel children-related bilingual dataset for emotion detection composed of two languages; (2) it evaluates the bilingual fine-grained emotion detection task and establishes strong baselines based on GPT and variants; (3) it exams different supervised and unsupervised pre-training techniques, shedding light on the significance of selecting the appropriate pre-training domain; (4) it addresses specific cross-cultural challenges in bilingual emotion detection tasks.

2 Literature review

2.1 Emotion analysis in linguistics

Along with the "Affective turn" (Kim & Bianco, 2007), which emphasizes the growing significance of affect as a focal point of analysis across various disciplinary and interdisciplinary discourses, research into emotion detection has gained substantial traction in the field of computational linguistics over the past two decades (Bilianos, 2022; Chen et al., 2009; Chuang & Wu, 2002; Lee et al., 2009; Lee et al., 2010; Mihalcea & Liu, 2006; Ogarkova, 2013; Peng et al., 2024; Picard, 1995/2000; Strapparava & Mihalcea, 2008). Current investigations in text-based emotion detection encompass various domains. One prominent avenue involves the examination of emotions within the context of online social media platforms. This encompasses a wide range of data sources, from theme-based book reviews and movie comments on platforms like Goodreads (Dimitrov et al., 2015) to the unfiltered expression of thoughts and sentiments on platforms such as Twitter and Reddit (Demszky et al., 2020). Another noteworthy direction centres around the analysis of emotions in literary classical works, including fairy tales (Mohammad, 2012), among others. However, it is essential to acknowledge that a substantial portion of research in this domain predominantly relies on monolingual datasets, which poses limitations in understanding the complexities of emotions in bilingual contexts. While some scholars have recognized the importance of incorporating bilingual or multilingual datasets, their approach often involves annotating emotions in a single language and subsequently relying on rudimentary translation methods, such as Google Translate, to render the annotated dataset to other languages (Mohammad & Turney, 2010). This methodology, unavoidably, introduces potential errors and inaccuracies in the resulting translated dataset. Therefore, a more dedicated exploration of the bilingual perspective is crucial to enhance emotion detection in Chinese-English language pairs, analyse emotional nuances in translations and cross-cultural studies, and improve cross-linguistic and cross-cultural analyses of emotions.

For the trends and application models in the identification and analyses of emotions in languages, there are mainly rule-based approach, machine learning-based approach and deep learning approach. In the rule-based emotion analysis, the use of emotion-bearing words and their combinations to assess phrasal units for emotions has been a primary focus of emotion analysis research for a long time (Aman & Szpakowicz, 2007; Chen et al., 2009; Lee et al., 2013). Popular emotion lexicons includes NRC Lexicon (Mohammad & Turney, 2010, 2013), ANEW (Bradley & Lang, 1999; Nielsen, 2011) and the Valence Arousal Dominance Lexicon (Mohammad, 2018). The machine learning-based approach to emotion analysis entails converting text emotion analysis into a classification task. This involves the application of established algorithms like Support Vector Machines (SVMs), Naive Bayes, Logistic Regression, and other machine learning methods (Aman & Szpakowicz, 2007; Danisman & Alpkocak, 2008; Deshpande & Rao, 2017). Although bag-of-words models have demonstrated promise in the domains of speech emotion recognition (Jain et al., 2020; Kwon et al., 2003) and facial emotion detection (Michel & El Kaliouby, 2003; Susskind et al., 2007), there exists considerable potential for refinement within the context of text-based emotion analysis in terms of sparse data features. In recent years, deep learning approach, such as Convolutional Neural Networks (CNNs), Long Short-Term Memory Networks (LSTMs), and Recurrent Neural Networks (RNNs), have gained significant prominence in the field of text emotion analysis (Ghosh et al., 2023a, 2023b; Zhou & Long, 2018). These approaches have garnered attention due to their demonstrated ability to address the inherent limitations associated with traditional machine learning techniques. More recent advancements in transformer-based models, such as BERT and GPT family, and DeepSeek, which incorporate language model pre-training, have demonstrated significantly improved performance (Guu et al., 2020; Zhang et al., 2024).

2.2 Taxonomy of emotion models

Previous research has predominantly employed diverse emotional taxonomies, including Ekman's (1992) six basic emotions (JOY, ANGER, FEAR, SADNESS, DISGUST, and SURPRISE), Panksepp's (2007) seven basic emotions (SEEKING, FEAR, RAGE, LUST, CARE, PANIC and PLAY), Plutchik's (2003) eight basic emotions (JOY, SADNESS, ANGER, FEAR, TRUST, DISGUST, SURPRISE, and ANTICIPATION), Parrott's (2001) five basic emotions (JOY, SADNESS, ANGER, FEAR, and LOVE), and the extensive GoEmotions taxonomy with over 27 emotions (Demszky et al., 2020). Besides these single-label classification systems, Izard's (1997) Differential Emotions Theory introduced intensity-graded dual labels (e.g., INTEREST-EXCITEMENT, ENJOYMENT-JOY, ANGER-RAGE, FEAR-TERROR). The emotion taxonomy in this dataset is founded on established research, acknowledging four primary emotions are HAPPINESS, SADNESS, ANGER, and FEAR (Lee, 2015).

This study adopts Parrott's (2001) classification, which includes LOVE except the four primary ones, aligning with the focus on children's literature. This adoption stems from two key considerations. Firstly, LOVE holds a pivotal role in children's emotional development. It serves as a cornerstone for nurturing empathy, fostering healthy attachment, and facilitating social growth (Haslip et al., 2019). Through experiences of LOVE and affection, children cultivate a sense of security, belonging, and emotional well-being, all essential elements for their overall resilience and development. Secondly, LOVE emerges as a consistent and significant theme in children's literature (Rustin & Rustin, 2018). Within these narratives, emotional themes are carefully crafted to resonate with young readers, aiming to evoke empathy and foster emotional engagement. By exploring the theme of LOVE within children's literature, we not only enrich the emotional landscape of our analysis but also contribute to reinforcing positive attitudes among children, thereby promoting the development of their emotional competence.

2.3 Emotion analysis in children's literature

Scholarly interest in the analysis of emotions in children's literature emerged as early as the late twentieth century, as evidenced by Stevenson (1997). Previous research in this field has primarily focused on several key areas, including the detection and recognition of emotions within children's literary texts, as explored by Alm and Sproat (2005), Alm et al. (2005) and more recently by Bizzoni et al. (2022), Herrmann (2023), Jularić (2020) and Zad et al. (2021). Additionally, scholars have examined the socio-cultural implications of emotions depicted in children's literature, as demonstrated by Adukia et al. (2022) research in 2022, and have conducted psychological examinations of hypotheses related to emotions, exemplified by Jacobs et al. (2020).

The methodologies for analysing emotions in children's literature align with broader emotion analysis techniques, utilizing both rule-based and machine learning approaches. For example, Saif Mohammad's studies in 2010 and 2013 examined emotions in novels and fairy tales, demonstrating the effectiveness of sentiment analysis combined with visualization for quantifying emotions in various texts. Alm et al. (2005) used the SNoW architecture for supervised machine learning to gain insights into text-based sentiment prediction. Additionally, Jacobs et al. (2020) focused on the Pollyanna Effect, employing the SentiArt tool for nuanced sentiment dynamics analysis, contributing significantly to the discourse on emotion in literature. While these studies share common methodological approaches, the specific focuses differ. For example, Herrmann (2023) primarily focused on developing and applying new measures, such as Average Valence, Emotional Potential, Emotional Arc, and Emotion Profile, to analyse a dedicated fairy tale corpus called ChildTale-A. In contrast, Bizzoni et al. (2022) investigated the potential link between sentiment development and perceived literary quality, using the Hurst exponent to assess the internal coherence and predictability of the sentiment arcs within the fairy tales of H.C. Andersen.

While transformer-based large language models (LLMs) – including BERT, the GPT family, and Chinese-specific models like DeepSeek and Qwen – have demonstrated remarkable success across diverse domains such as legal language analysis (Li et al., 2025), tourism applications (Xu et al., 2024), and psychological text processing (Rathje et al., 2024), their potential for emotion analysis in children's literature remains largely unexplored. To address this gap in the literature, the current study aims to leverage BERT, GPT family, DeepSeek and Qwen2.5 models in conjunction with high-quality data annotated with emotions. This approach seeks to enhance our understanding of emotional content within children's literature, drawing upon the latest developments in natural language processing techniques.

3 Dataset

3.1 BilingualChildEmo

This paper presents BilingualChildEmo, a non-parallel bilingual Chinese-English emotion dataset comprising 2,177 sentences (53.4% Chinese, 46.6% English)[1] annotated with five basic emotions (JOY, SADNESS, ANGER, FEAR, and LOVE). The dataset was sampled from a parallel children's literature corpus of approximately 50,000 aligned sentence pairs, which includes original English texts and their Chinese translations in different periods across multiple genres (see Appendix 1). Crucially, to enable authentic cross-linguistic emotion analysis and avoid translation bias, we employed a non-parallel sampling strategy: 1162 sentences were randomly selected from the Chinese translations and 1015 from the English originals (representing 25% of total selections from each language pool), ensuring no direct translation pairs were included. This approach preserves the corpus's historical and linguistic diversity while creating balanced, independent samples that reflect native emotional expressions in each language. For reference, Table 1 displays representative examples from the BilingualChildEmo dataset.

Our adoption of Parrott's (2001) taxonomy—which notably includes LOVE alongside the four primary emotions (JOY, SADNESS, ANGER, and FEAR)—receives both theoretical (see discussion in Sect. 2.2) and empirical support, as the correlation analysis (Table 2) validates its appropriateness for children's literature. The results shown in Table 2 indicate that all emotion categories exhibit negative correlations with one another, suggesting that the adopted emotion categories are distinct within the dataset. Notably, there is a strong negative correlation (– 0.427) between JOY and SADNESS, highlighting their extensive mutual exclusivity. However, the weaker correlations observed among ANGER*LOVE (– 0.141),

[1] While not perfectly 50%-50% balanced, this distribution (53.4% Chinese, 46.6% English) was intentionally maintained to preserve the natural prevalence of emotional expressions found in our source materials, rather than imposing artificial parity that might distort ecological validity.

Table 1 Examples in the BilingualChildEmo dataset

Sample Text	Label
So overjoyed were they at their deliverance that they laughed aloud, and the Earth seemed to them like a flower of silver, and the Moon like a flower of gold	JOY
巨人欣喜若狂地跑下楼梯, 出了房子冲进花园。	
And in the morning he rose up, and plucked some bitter berries from the trees and ate them, and took his way through the great wood, weeping sorely	SADNESS
可怜的人儿, 失去了他们唯一的儿子!	
织工气愤地看着他, 说: 你看我干什么?	ANGER
"Upon my word," said the Miller with anger, "you are very lazy."	
But his face was strangely pale, and as he fell upon the deck the blood gushed from his ears and nostrils	FEAR
据说那个墓穴里还躺着一个人, 死者是一个异常英俊美丽的青年, 他的双手用绳子反绑着, 胸部被捅了很多刀, 衣服都被血染红了。	
I am his best friend, and I will always watch over him, and see that he is not led into any temptations	LOVE
比如说, 新娘和新郎这么年轻就彼此相爱了。	

Table 2 Dataset correlation analysis

		JOY	SADNESS	ANGER	FEAR	LOVE
JOY	Pearson Correlation	1	−0.427**	−0.297**	−0.301**	−0.247**
	Sig. (2-tailed)		0.000	0.000	0.000	0.000
	N	2177	2177	2177	2177	2177
SADNESS	Pearson Correlation	−0.427**	1	−0.244**	−0.247**	−0.203**
	Sig. (2-tailed)	0.000		0.000	0.000	0.000
	N	2177	2177	2177	2177	2177
ANGER	Pearson Correlation	−0.297**	−0.244**	1	−0.172**	−0.141**
	Sig. (2-tailed)	0.000	0.000		0.000	0.000
	N	2177	2177	2177	2177	2177
FEAR	Pearson Correlation	−0.301**	−0.247**	−0.172**	1	−0.143**
	Sig. (2-tailed)	0.000	0.000	0.000		0.000
	N	2177	2177	2177	2177	2177
LOVE	Pearson Correlation	−0.247**	−0.203**	−0.141**	−0.143**	1
	Sig. (2-tailed)	0.000	0.000	0.000	0.000	
	N	2177	2177	2177	2177	2177

**. Correlation is significant at the 0.01 level (2-tailed)

FEAR*LOVE (−0.143), and ANGER*FEAR (−0.172) may indicate more complex relationships that warrant further investigation.

Furthermore, it is essential to elucidate the rationale behind adopting a sentence-level approach for emotion detection within this study. The decision to focus on sentences as the fundamental unit of analysis is underpinned by the belief that sentences represent a more contextually appropriate and semantically meaningful unit for the expression of emotions within the literary genre (Yang & Cardie, 2014). While this approach acknowledges, to some extent, the potential for complex linguistic devices such as irony and metaphor, which can be challenging to discern when examining emotions solely at the word level, it is worth noting that detecting irony and metaphors remains challenging even at the sentence level. This suggests that future studies may benefit from providing additional context to accurately identify figurative expressions.

3.2 Annotation

Six annotators were recruited for this dataset annotation: three native English speakers for labelling English data and three native Chinese speakers for labelling Chinese data. All annotators were postgraduate students with academic backgrounds in childhood studies and intercultural communication, ensuring their familiarity with both linguistic and developmental aspects of emotion recognition in texts. Prior to the annotation task, they received comprehensive training from a language expert, who reviewed the emotion taxonomy, provided detailed guidelines (including positive and negative identification rules, see Appendix 2), and addressed any questions to ensure consistent understanding.

Annotators were instructed to select a singular emotion descriptor for each sentence, opting for the emotion that they felt most confident in attributing. They were encouraged to consider both explicit expressions and the implicit contextual inference to identify the emotion conveyed in the sentence. Emotion labels were to be assigned based on the predominant emotion conveyed by the sentence, with annotators prioritizing the most dominant emotion in cases of multiple emotions. Annotations were expected to be consistent with the definitions provided above, taking into account cultural nuances and linguistic expressions. The expert reviewed all annotations to ensure adherence to the established guidelines, refining ambiguous cases through iterative consensus.

In instances where annotators encountered difficulties in annotating a particular sentence, such as when emotions were not readily discernible or when overly intricate emotions were present, they were guided to employ the label 'Other' for annotation purposes. For instance, in Example 1, emotions were not readily discernible, while in Example 2, emotions were deemed rather complicated and overwhelming. Sentences designated as 'Other' were subsequently excluded from the dataset during the compilation process.

Example 1: 小燕子開始想事，很快就入睡了。(Back translation: Little Swallow began to ponder and soon fell asleep.) (Emotion: Other).

Example 2: The child of the old King's only daughter by a secret marriage with one much beneath her in station—a stranger, some said, who, by the wonderful magic of his lute-playing, had made the young Princess love him; while others spoke of an artist from Rimini, to whom the Princess had shown much, perhaps too much honour, and who had suddenly disappeared from the city, leaving his work in the

Table 3 Results of Fleiss Multirater Kappa (overall agreement)

	Kappa	Asymptotic			Asymptotic 95% confidence interval	
		Standard error	z	Sig	Lower bound	Upper bound
Overall agreement	0.697	0.007	105.640	0.000	0.696	0.697

Table 4 Results of Fleiss Multirater Kappa (agreement on individual categories)

Rating category	Conditional probability	Kappa	Asymptotic			Asymptotic 95% confidence interval	
			Standard error	z	Sig	Lower bound	Upper Bound
JOY	0.335	0.764	0.012	61.768	0.000	0.764	0.765
SADNESS	0.265	0.694	0.012	56.077	0.000	0.693	0.695
ANGER	0.142	0.709	0.012	57.324	0.000	0.709	0.710
FEAR	0.152	0.675	0.012	54.569	0.000	0.674	0.676
LOVE	0.107	0.558	0.012	45.120	0.000	0.558	0.559

Cathedral unfinished—he had been, when but a week old, stolen away from his mother's side, as she slept, and given into the charge of a common peasant and his wife, who were without children of their own, and lived in a remote part of the forest, more than a day's ride from the town. (Emotion: Other).

To gauge the level of agreement among the annotators, agreement scores were computed both for overall agreement and individual emotions, utilizing Fleiss' Multirater Kappa statistic, as detailed in Tables 3 and 4. The computed average Fleiss' kappa value across all emotion categories was 0.697, with individual values ranging from 0.558 to 0.764. Following the interpretative framework elucidated by Landis and Koch (1977), kappa values exceeding 0 signify varying degrees of agreement that surpass mere chance among two or more raters, with a maximum attainable value of + 1 denoting perfect agreement, indicating complete consensus among the raters on all items. The overall agreement score suggests that the annotators shared a reasonable level of consistency in identifying emotions across the dataset. However, the variability in individual category scores highlights the need for ongoing refinement in emotion classification to enhance clarity and accuracy.

3.3 Data analysis

3.3.1 Emotion distribution

Figure 1 visually represents a scatter plot, offering insights into the distribution of the five primary emotions within the BilingualChildEmo dataset. This dataset

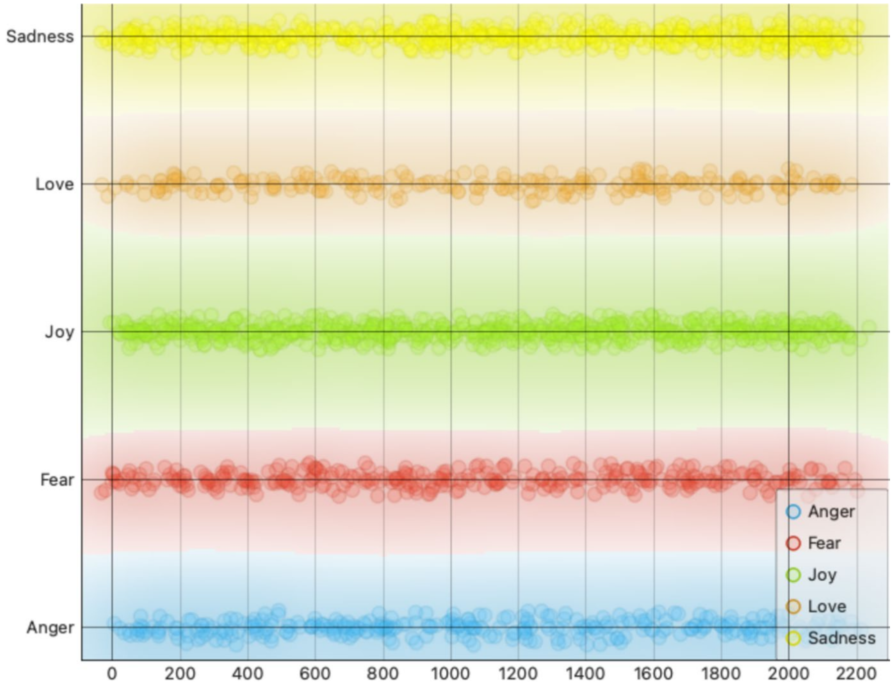

Fig. 1 Scatter plot of all the sentences

encompasses a total of 2,177 instances of emotions, with JOY emerging as the most prevalent emotion, occurring 745 times. Overall, the patterns are in line with findings in other or general domains. In close proximity, SADNESS follows closely with 565 occurrences. Existing scholarship, exemplified by Nikolajeva (2013), has posited that children's literature frequently incorporates themes centred around JOY and SADNESS. These emotionally vivid and contrasting themes are believed to aid children in comprehending and managing their own "emotion literacy" (p. 249). The substantial prevalence of JOY and SADNESS within the corpus may be indicative of a deliberate emphasis on these emotional themes within the children's literature contained in this dataset.

Conversely, ANGER and FEAR are observed less frequently, appearing 315 and 323 times, respectively. This lower incidence may be reflective of the nature of children's literature, as noted by Logan (1998), which typically avoids explicit depictions of violence or frightening content to protect young readers' emotional well-being. Lastly, LOVE is the least frequently depicted emotion, with only 229 instances. This infrequency may stem from the complexity of LOVE itself, as young readers might struggle to grasp its nuances. The dataset correlation analysis in Table 2 reinforces this idea by showing that, while LOVE is recognized as a distinct emotional category, it exhibits weaker negative correlations (ranging from − 0.141 to − 0.247) with other emotions. Additionally, the lowest agreement score among annotators for the LOVE category (0.558) indicates the challenges they face in consistently identifying this subjective and context-dependent emotion. Together, these factors suggest that the intricacies of LOVE contribute to its limited portrayal in the dataset.

Overall, the distribution of emotions within the BilingualChildEmo dataset, as illustrated in Fig. 1, reveals a relatively balanced profile, despite notable variations in the proportional representation of each emotion. The findings, however, are dataset-specific and require further investigation to confirm their broader applicability to children's literature.

3.3.2 A comparison of emotions in Chinese and English sentences

Figure 2 provided emotion distribution data, presenting an intriguing comparison between the English and Chinese language samples included in the BilingualChildEmo dataset. While we cannot extrapolate these findings to draw conclusions about the emotional landscapes of the English and Chinese-speaking populations as a whole, the data does offer valuable insights into the nuances within the given dataset. Within the confines of this dataset, the higher proportion of JOY observed in the English samples (36.06%) compared to the Chinese samples (32.36%) suggests that positive emotional expressions may be more prevalent in the English language data under examination. Similarly, the greater incidence of SADNESS in the Chinese samples (28.23%) relative to the English samples (23.55%) indicates that this particular emotional state may be more strongly represented in the Chinese language data within the dataset.

The slightly higher percentages of ANGER in the English data (15.47%) versus the Chinese data (13.60%), as well as the marginally greater presence of FEAR in the Chinese data (15.49%) compared to the English data (14.68%), point to

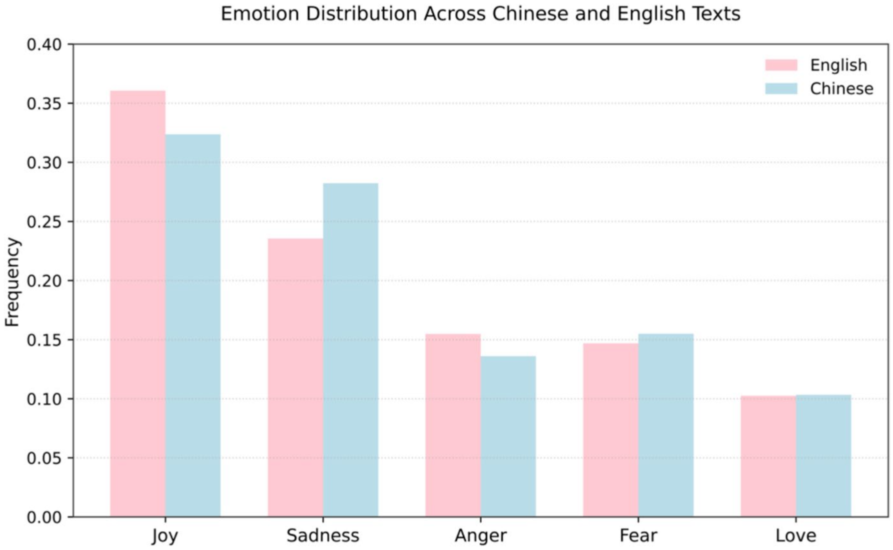

Fig. 2 Emotion distribution across Chinese and English texts

potential differences in the distribution of these emotional categories within the confines of this specific dataset. This observation further underscores the influence of culture on emotional themes within literature. For instance, in Chinese culture, sentences associated with supernatural elements are often categorized as inducing FEAR (See “耶穌Jesus” in Example 3).While in English, such expressions are typically considered neutral and lack the connotations of horror (See “Christ” in Example 4) commonly found in translated Chinese literature.

Example 3: 他跪在耶穌像前, 神龕旁的大蠟燭燃燒得很亮, 香燒起的縷縷青煙在穹蓋形成薄薄的霧環。(Back translation: He knelt in front of the statue of Jesus. The big candle next to the shrine burned brightly, and the wisps of green smoke from the burning incense formed a thin ring of mist on the dome.) (Emotion: FEAR).

Example 4: He knelt before the image of Christ, and the great candles burned brightly by the jewelled shrine, and the smoke of the incense curled in thin blue wreaths through the dome. (Emotion: JOY).

Additionally, the portrayal of the motif of death in the current dataset of children’s stories highlights cultural contrasts among Chinese and English. In Example 5, the English text associates the presence of a dead bird with SADNESS, while the Chinese translation in Example 6 conveys a sense of FEAR. This disparity may stem from differing cultural attitudes toward death—where Chinese culture may treats it as a taboo subject (Hsu et al., 2009), Western cultures may view it as a natural aspect of life (Palgi & Abramovitch, 1984).

Example 5: ‘And here is actually a dead bird at his feet!’ continued the Mayor. (Emotion: SADNESS).

Example 6: “他脚边还有一只死鸟!”市长又说。(Back Translation: ‘And here is actually a dead bird at his feet!’ continued the Mayor.) (Emotion: FEAR).

The near-identical proportions of LOVE between the English (10.25%) and Chinese (10.33%) samples suggest that this particular emotion may be expressed with similar prevalence within the dataset, potentially indicating universal or cross-cultural similarities in the conceptualization and expression of LOVE. Yet, even this observation should be tempered, as it may be specific to the particular dataset and not necessarily reflective of the broader language and cultural landscapes.

In conclusion, the provided emotion distribution data offers a valuable, yet limited, glimpse into the nuanced differences in emotional expression between the English and Chinese language samples within this specific dataset. These findings highlight the potential for translations to evoke changes in the portrayal and perception of emotions, illuminating the unique characteristics of translated literature. For instance, in Example 7, the translated Chinese sentence “這絕唱 (This final song)” conveys the emotion of SADNESS, while its source text “It” remains emotionally neutral. This exemplifies how translations may imbue target language sentences with additional emotional nuances not present in the source text.

Example 7: 這絕唱隨著河流的浪頭飄去， 把它的餘音一直傳向大海。 (Back translation: This final song floats away with the waves of the river, continuously transmitting its lingering sound towards the sea.) (Emotion: SADNESS).

Source text: It floated through the reeds of the river, and they carried its message to the sea. (Note: this sentence is just the source text and is not included in the BilingualChildEmo dataset so it's not labelled with emotions).

However, it is important to note that this data represents a limited dataset of children's literature and should not be used to draw broad generalizations about the emotional landscapes of the English and Chinese-speaking societies as a whole. Further research and validation across multiple datasets and contexts would be necessary to gain a more comprehensive understanding of the complex interplay between language, culture, and the expression of emotions, particularly in the realm of translated literature.

4 Modelling methods

This study models Parrot's core set of five emotions within the BilingualChildEmo dataset using a range of machine learning and deep learning approaches. These include supervised learning techniques like Support Vector Machines (SVMs), Random Forests, and Logistic Regression; unsupervised learning through Neural Networks; and self-supervised learning by fine-tuning transformer models such as XLM-RoBERTa, the GPT family, DeepSeek, and Qwen.

Initially, the data preparation phase begins with frequency calculations for both the English source text, using AntConc, and the translated Chinese target texts, employing Sketch Engine. The data is then tokenized for emotion analysis through the Stanford NLP Group's Stanza. Following this, the dataset is partitioned into training, validation, and testing sets, comprising 70%, 15%, and 15% of the data, respectively. Both supervised and unsupervised learning models are evaluated, alongside Multilingual Sentence-BERT (SBERT) word embeddings (Reimers & Gurevych, 2019), as the chosen methodology.

The analysis progresses with the application of various techniques, starting with supervised machine learning methods (SVMs, Random Forest, and Logistic Regression), followed by unsupervised deep learning using Neural Networks. The process culminates in fine-tuning advanced transformer models, specifically XLM-RoBERTa, several GPT variants (gpt-4o-2024-08-06, gpt-4o-mini-2024-07-18), DeepSeek variant (DeepSeek-R1-Distill-Llama-8B), and Qwen2.5-7B on annotated datasets. To address the imbalanced dataset, we experimented with and adjusted the training parameters, particularly class weights, to identify the optimal settings for SVMs, Random Forest, and Logistic Regression (see Appendix 3). For XLM-RoBERTa, DeepSeek variant, and Qwen2.5-7B, the training parameters comprise a batch size of 3, a total of 3 training epochs, and a learning rate of 2e-5. The Transformer finetuning models are configured with a batch size of 3, 3 training epochs, and a learning rate multiplier of 1.8. All the models underwent comprehensive fine-tuning of their entire model architecture to optimize performance on the task at hand.

Model performance is evaluated using a range of metrics, including F1 score, precision, recall, and accuracy, which collectively demonstrate promising results in emotion prediction. For F1 metric, we report three variations of the F1 score: F1-micro, F1-macro, and a weighted F1 score. The F1-micro provides an overall measure of the model's performance by calculating the F1 score across all emotion categories, treating them equally regardless of their frequency. The F1-macro, on the other hand, calculates the F1 score for each emotion category individually and then takes the average, giving equal weight to each category. Finally, the weighted F1 score takes into account the class imbalance in the dataset, weighting the F1 score of each category proportionally to its prevalence. The schematic representation of the workflow is depicted in Fig. 3 from Orange platform, providing a visual synopsis of the entire process.

Additionally, the study investigates transformer-based language models, particularly XLM-RoBERTa, GPT and DeepSeek, renowned for their excellence in natural language processing tasks (Lauriola et al., 2022; Worsham & Kalita, 2020). The primary objective is to further elevate the accuracy of emotion classification. To attain this goal, the study initiates with a comprehensive examination of the structural intricacies and algorithmic foundations characterizing transformer-based language models. Subsequently, the training phase unfolds, employing the comprehensive BilingualChildEmo dataset, which encapsulates a diverse array of emotional expressions within the context of children's literature, spanning a rich tapestry of linguistic styles and contextual nuances. Following the rigorous training regimen, the model is subjected to a meticulous evaluation, wherein a comprehensive array of performance metrics, including the F1 score and accuracy, is invoked to methodically assess its efficacy and robustness.

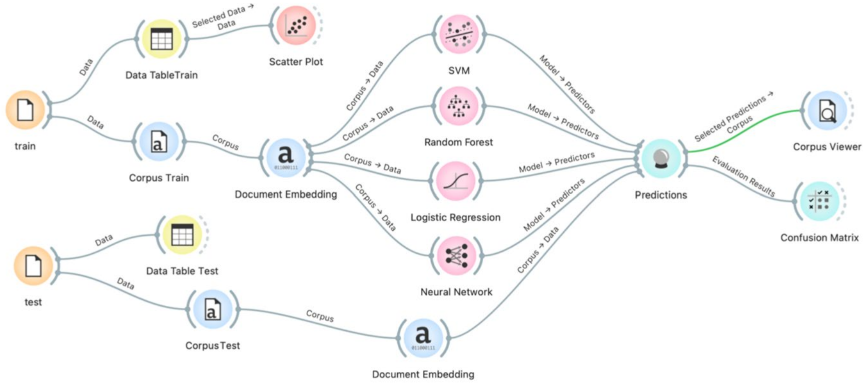

Fig. 3 Workflow of modelling methods

4.1 Supervised machine learning: SVMs, random forest and logistic regression

This study initially focuses on three prominent supervised machine learning algorithms—SVMs, Random Forest, and Logistic Regression—for the task of emotion classification. SVMs is recognized for its efficacy in constructing optimal hyperplanes to delineate data points into distinct classes, thereby maximizing inter-class separation and ensuring robust classification performance (Raschka & Mirjalili, 2019). Within the scope of this research, SVMs is employed to discern various emotional categories by learning discernible patterns within the provided training data. Random Forest operates as an ensemble method that combines multiple decision trees through bagging, demonstrating exceptional capability in handling high-dimensional and imbalanced data by leveraging feature subsampling and majority voting. In emotion classification, Random Forest excels at capturing complex non-linear relationships between linguistic features—such as lexical choices and syntactic patterns—and emotional categories. Logistic Regression, a linear model well-suited for binary or multiclass classification tasks, estimates the probability of data points belonging to specific classes by fitting a logistic function to the input features (Sen et al., 2020). In the context of this study, Logistic Regression is harnessed to forecast the probability of an instance being associated with a particular emotion, grounded in its acquired knowledge of the relationships between features and emotional categories.

4.2 Unsupervised deep learning: neural network

Differing from supervised machine learning, unsupervised deep learning is directed at uncovering latent patterns and structures within unlabelled data (Raschka & Mirjalili, 2019). Neural networks, a category of deep learning models, exhibit remarkable versatility and can be applied to unsupervised learning tasks, including the domain of emotion classification. These computational models draw inspiration from the human brain and consist of interconnected layers of artificial neurons (Batool et al., 2013). They possess the capacity to acquire intricate patterns and representations from raw data, rendering them apt for a multitude of applications, spanning image and text classification. In unsupervised learning contexts, neural networks prove instrumental in the identification of inherent data structures, such as clusters or low-dimensional representations, which subsequently find utility in emotion classification endeavours.

4.3 Transformers with LLMs: XLM-RoBERTa, GPT, DeepSeek and Qwen

In recent years, self-supervised learning methodologies have garnered considerable attention due to their aptitude for harnessing substantial volumes of unannotated data to pretrain models, subsequently amenable to fine-tuning for task-specific objectives (Atito et al., 2021). This study directs its attention towards transformer models, specifically emphasizing (1) XLM-RoBERTa and (2) members of the GPT family, encompassing gpt-3.5-turbo, gpt-4o, gpt-4o-mini, and (3) Chinese large language models such as DeepSeek and Qwen, as instruments for the task of emotion classification.

XLM-RoBERTa, an advanced transformer-based language model developed by Facebook AI, extends upon the achievements of the RoBERTa model. It is pretrained on a vast multilingual corpus spanning over 100 languages, enabling it to capture the intricacies of diverse linguistic structures. Several factors contribute to the selection of XLM-RoBERTa for this study. Firstly, its pretraining includes text sources that closely resemble children's literature, such as book corpora and stories, which aligns well with the focus of the current study. Secondly, its multilingual capabilities broaden its applicability, particularly in the context of language diversity. Lastly, the model employs Sentence-BERT (SBERT) (Reimers & Gurevych, 2019) for sentence-level embeddings, aligning with the study's objectives and providing a robust foundation for research. Consequently, this study will undertake fine-tuning procedures on the XLM-RoBERTa model.[2]

The GPT family of language models, including gpt-3.5-turbo, gpt-4o, and gpt-4o-mini, have also gained significant attention in recent years. Developed by OpenAI, these models have demonstrated impressive performance across a wide range of natural language processing tasks (Min et al., 2023; Naveed et al., 2023). Their strong generative capabilities and ability to capture contextual information make them promising candidates for emotion classification as well. This study will investigate the efficacy of fine-tuning various GPT models for the given task, exploring their potential to complement or outperform the XLM-RoBERTa approach.

Chinese large language models, such as DeepSeek and Qwen, may represent a promising area of exploration in this study. With the increasing demand for emotion classification in non-English contexts, these models offer distinct advantages through their specialized training on Chinese datasets. DeepSeek demonstrates exceptional capabilities in contextual understanding and nuanced language generation, providing a valuable resource for analysing emotional subtleties in Chinese text (Guo et al., 2025). Similarly, Qwen is specifically designed to navigate various linguistic challenges inherent in the Chinese language, making it a robust tool for emotion classification tasks. Our selection of these models aligns with the purpose of this study, as our bilingual dataset is specifically crafted to leverage the strengths of these advanced tools, enabling a comprehensive analysis of emotional expressions in both Chinese and English contexts.

[2] The finetunded model based on XLM-RoBERTa is available on HuggingFace: https://doi.org/10.57967/hf/1912.

Table 5 The supervised and unsupervised classification experiment results by emotion

Model	JOY		SADNESS		ANGER		FEAR		LOVE	
	F1	CA	F1	CA	F1	CA	F1	CA	F1	CA
Logistic Regression	0.781	0.726	0.571	0.513	0.583	0.600	0.550	0.574	0.587	0.710
Neural Network	0.763	0.748	0.590	0.571	0.568	0.651	0.561	0.549	0.614	0.548
Random Forest	0.733	0.832	0.600	0.536	0.462	0.419	0.559	0.314	0.609	0.548
SVMs	0.749	0.766	0.616	0.583	0.517	0.535	0.538	0.549	0.602	0.595

The best performing model F1s and CA are underlined and highlighted

5 Experiment results

5.1 Supervised and unsupervised classification results

Table 5 presents the performance of four classification models—Logistic Regression, Neural Networks, Random Forest, and SVMs—in predicting five emotion categories: JOY, SADNESS, ANGER, FEAR, and LOVE. Evaluation metrics include F1 scores and classification accuracy (CA). The results indicate that Logistic Regression and Neural Network consistently outperform other models, achieving high F1s and CAs across multiple categories, while SVMs follow closely with a high F1 score for SADNESS of 0.616. Conversely, Random Forest achieves a notably high CA for JOY but comparatively lower scores for other emotions.

Figure 4 visually compares the F1-score performance of four classification models across five emotional categories. The results demonstrate that JOY achieves consistently superior classification (F1: 0.733–0.781), suggesting its expressions contain the most distinctive linguistic markers. In contrast, ANGER displays the greatest performance variability (ΔF1 = 0.121 across models), revealing fundamental challenges in detecting anger expressions. While SADNESS maintains stable classification (F1: 0.571–0.616) with SVMs performing best (F1 = 0.616), FEAR proves most difficult to classify (max F1 = 0.561), reflecting its contextual complexity. LOVE shows moderate but consistent recognition (F1: 0.587–0.614), with Neural Networks achieving peak performance (F1 = 0.614), indicating relatively stable

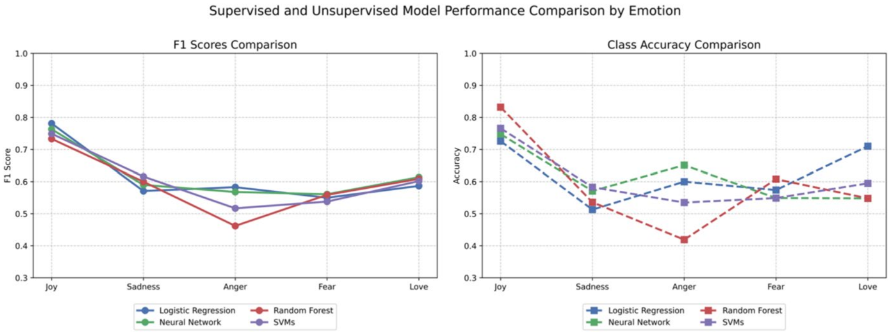

Fig. 4 Supervised and unsupervised classification F1 results

Table 6 The average supervised and unsupervised classification experiment results

Model	F1 Micro	F1 Macro	F1 Weighted	CA
Logistic Regression	0.640	0.639	0.615	0.640
Neural Network	0.642	0.642	0.619	0.642
Random Forest	0.630	0.598	0.624	0.630
SVMs	0.633	0.605	0.633	0.633

The best performing model F1s and CA are underlined and highlighted

but still challenging detection of affectionate language. These findings collectively emphasize the necessity of emotion-specific modelling strategies in affective computing systems, particularly for nuanced emotions like FEAR and ANGER that require specialized feature extraction approaches.

Table 6 summarizes the average results of supervised and unsupervised classification experiments across various models, measured by F1 Micro, F1 Macro, F1 Weighted, and Classification Accuracy. The results indicate that Neural Networks excel in emotion classification, achieving the highest F1 Micro/Macro score (0.642) and Classification Accuracy (0.642). SVMs follow closely with consistent scores (F1 Micro/Weighted/CA all at 0.633), and they exhibit a noticeable drop in F1 Macro (0.605), suggesting slightly weaker performance on minority classes. Logistic Regression maintained stable performance (F1 Micro/Accuracy: 0.640) but struggled with class weighting (0.615). Random Forest exhibited the largest metric variance, with an F1 Macro of 0.598 versus an F1 Weighted score of 0.624, reflecting its sensitivity to class distribution. The consistent superiority of Neural Networks across F1 Macro and F1 Weighted metrics highlights the advantages of advanced architectures in this domain.

One noteworthy facet to consider when interpreting the findings of this study pertains to the influence of feature extraction techniques on model efficacy. In this experiment, all models were tested with SBERT word embeddings, which capture semantic information at the sentence level. This approach allows the models to incorporate valuable context when making predictions (Reimers & Gurevych, 2019), which is crucial for understanding and predicting emotions in text. However, it is worth noting that different feature extraction techniques might yield different results, and future studies could explore alternative methods, such as bag-of-words, term frequency-inverse document frequency (TF-IDF), or word2vec, to compare their effectiveness in the context of emotion classification.

5.2 Transformer-based classification results

This study also delves into the prospective capabilities of transformer-based language models, including XLM-RoBERTa, GPT-4o, GPT-4o-mini, Qwen-2.5-7B, and DeepSeek-R1, to enhance the accuracy of emotion classification through

Table 7 Transformer-based model classification experiment results

Model	F1 Micro	F1 Macro	F1 Weighted	CA
XLM-RoBERTa	0.702	0.667	0.706	0.702
gpt-4o-2024-08-06	<u>0.779</u>	<u>0.764</u>	<u>0.778</u>	<u>0.779</u>
gpt-4o-mini-2024-07-18	0.752	0.741	0.754	0.752
Qwen-2.5-7B	0.646	0.660	0.601	0.646
DeepSeek-R1-Distill-Llama-8B	0.724	0.715	0.715	0.724

The best performing model F1s and CA are underlined and highlighted

Table 8 Overall agreement metrics

	Kappa	Asymptotic			Asymptotic 95% confidence interval	
		Standard error	z	Sig	Lower bound	Upper bound
Overall Agreement (3 Annotators)	0.643	0.017	38.492	0.000	0.642	0.644
Overall Agreement (2 Annotators + 1 Model Prediction)	0.479	0.017	28.337	0.000	0.478	0.480

multiclass categorization. Table 7 provides an overview of the multiclass classification results obtained via fine-tuning these four models. The different F1 scores and classification accuracy are meticulously documented for each epoch of the training process.

The results reveal significant performance differentiation among transformer models, with GPT-4o-2024-08-06 dominating across all metrics (F1 Micro: 0.779, F1 Weighted: 0.778, CA: 0.779), reflecting its superior architectural capacity to capture nuanced emotional patterns. Its F1 Macro score (0.764) further confirms balanced performance across all emotion classes. The GPT-4o-mini-2024-07-18 achieves competitive results (F1 Micro: 0.752, CA: 0.752) with reduced parameters, demonstrating efficient optimization for resource-constrained deployments. Among models with Chinese language specialization, DeepSeek-R1 delivers competent results (F1 Micro: 0.724), while Qwen-2.5-7B shows particular challenges with class-imbalanced distributions (F1 Weighted: 0.601 vs Macro: 0.660). XLM-RoBERTa's multilingual architecture achieves intermediate results (F1 Micro: 0.702), outperforming the Chinese-specialized models while still lagging behind GPT-4 variants. These findings show that while larger models (e.g., GPT-4o) excel in generalization, optimized smaller variants (e.g., GPT-4o-mini) remain viable for practical applications, and architectural choices significantly impact performance on nuanced tasks like emotion classification.

6 Model evaluation and error analysis

To evaluate model performance, we adopted a method where the label of one annotator is removed from the test set and replaced with the model's prediction. This allows for a comparative analysis of Kappa statistics between two annotator labels and one model prediction versus the agreement among all three annotators. The results presented in Table 8 highlight a substantial level of agreement (Kappa = 0.643) among the three annotators, indicating a high degree of consistency in their assessments. In contrast, the agreement between two annotators and the model's prediction yielded a lower Kappa value of 0.479, suggesting that while the model performs reasonably well, there is still room for improvement in aligning with human judgement.

The inter-rater reliability analysis (Table 9) reveals systematic differences in emotion classification between human annotators and the model. Human annotators achieved substantial agreement for JOY (κ = 0.700) and SADNESS (κ = 0.655), while the model showed significantly lower agreement for these categories (JOY: κ = 0.468; SADNESS: κ = 0.462). This pattern is exemplified in Example 8, where the Chinese sentence "她从自己的腰带里抽出一把绿蛭蛇皮柄小刀，递给他" ("She drew a green-leech-snakeskin-hilted dagger from her belt and handed it to him") produced divergent human annotations (2 × FEAR, 1 × LOVE) versus the model's uniform FEAR prediction. This case demonstrates how human judgment (annotator 2) incorporates complex social and cultural knowledge (interpreting weapon-giving between a "she" and "him" as potentially romantic), whereas the model relies primarily on lexical threat cues ("snake", "knife"). The moderate model performance on LOVE (κ = 0.529) and FEAR (κ = 0.504) suggests current architectures can recognize concrete emotion indicators but struggle with contextual and cultural complexities, particularly for ANGER (κ = 0.465). These findings highlight the need for more sophisticated context-aware modelling approaches that better capture the pragmatic and sociocultural dimensions of emotional expression.

Example 8: 她从自己的腰带里抽出一把绿蛭蛇皮柄小刀，递给他。(Back Translation: She drew a small green leech snakeskin-handled knife from her own belt and handed it to him.) (Annotator 1: Fear; Annotator 2: Love; Annotator 3: Fear; model prediction: Fear).

To complement the Kappa analysis, the error analysis in Fig. 5 offers additional insights into the model's error patterns. This would help identify specific

Table 9 Agreement on individual emotion categories

	JOY	SADNESS	ANGER	FEAR	LOVE
Kappa (3 Annotators)	0.700	0.655	0.631	0.649	0.533
Kappa (2 Annotators + 1 Model Prediction)	0.468	0.462	0.465	0.504	0.529

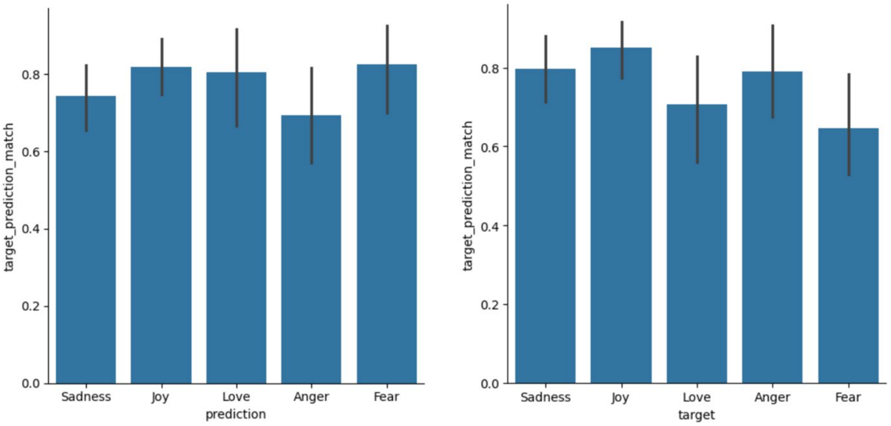

Fig. 5 GPT-4o Emotion Classification Error Analysis

misclassifications and guide targeted improvements in the model's ability to distinguish between emotions. The model demonstrates commendable accuracy in identifying JOY, achieving an accuracy of 85.05%, and performs relatively well with SADNESS, reaching an accuracy of 79.76%. These results indicate the model's effectiveness in recognizing positive emotional expressions. In contrast, the model exhibits notable difficulties with FEAR and ANGER, with accuracies of 64.71% and 79.07%, respectively. This discrepancy suggests that the model may require more nuanced feature representation or additional training data to better capture these complex emotional states.

The detailed breakdown of predictions provides further insights. For ANGER, the target accuracy is 79.07%, but the prediction accuracy is only 69.39%, indicating that the model tends to underpredict instances of ANGER, frequently misclassifying them as FEAR or JOY within particular contexts. This pattern suggests the model struggles to distinguish between emotions of similar arousal levels but differing valence (ANGER-FEAR) as well as opposite-valence emotions (ANGER-JOY). The frequent confusion between positive (JOY) and negative (ANGER, FEAR) valence categories indicates that incorporating sentiment polarity as an additional feature could possibly substantially improve classification. In contrast, for FEAR, the target accuracy is 64.71%, while the prediction accuracy is 82.50%. Although the model shows slight improvement in predicting FEAR, it still misclassifies it as ANGER or SADNESS on numerous occasions.

The confusion matrix for GPT-4o in Fig. 6 reveals significant misclassifications,[3] particularly where the true emotion label is FEAR but the model predicts SADNESS. This indicates the model has difficulty distinguishing between these two closely related emotions in the given context. For example, in cases like "One day the Giant came back." and "她的歌儿越来越细弱, 她觉得有什么东西堵住了她的

[3] The confusion matrixes for all the models are available in Appendix 4.

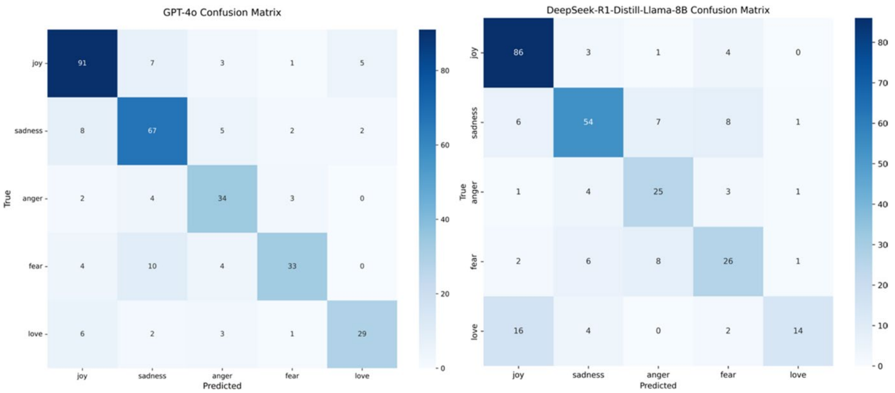

Fig. 6 Confusion Matrix for GPT-4o and DeepSeek-R1

喉咙。", the true emotion is FEAR but the model incorrectly predicts SADNESS. This firstly suggests that emotions like FEAR and SADNESS can be inherently complex and nuanced, and may co-occur or be challenging to differentiate based solely on sentence-level information. The challenges in accurately classifying these related emotions highlight the limitations of single-label emotion annotation. Future studies may consider exploring multi-label emotion assignments to better capture the subtle nuances present in the data.

Example 9: “呜呜呜!”狼嗥叫着, 两条后腿夹着尾巴慢吞吞地在灌木丛里穿行, “这真是冻死人的鬼天气。” (Back Translation: “Woo, woo, woo!” the wolf howled, moving slowly through the bushes with its tail between its hind legs, “This is truly ghost weather that freezes people to death.”) (Target: ANGER; Prediction: FEAR).

Another factor contributing to the model’s difficulties in accurately classifying certain emotional categories, such as FEAR and ANGER, is the language-specific and cultural differences between the primarily English-trained GPT models and the Chinese language context. For instance, in Example 9, the term “鬼” (ghost) is employed to express feelings of ANGER, illustrating a cultural nuance where this word serves as an intensifier, conveying a sense of severity rather than its literal association with ghosts. In Chinese, “鬼” (ghost) placed before a noun acts as a modifier, enhancing the degree of the emotion, functioning idiomatically rather than literally. Such subtle language-specific and culturally-bound emotional expressions can be challenging for machine learning models trained on more generalized, English-based data to accurately capture and interpret, leading to discrepancies in emotional classification.

In comparing the English-based LLM GPT-4o with the Chinese LLM DeepSeek-R1, it is clear that DeepSeek-R1 exhibits a higher rate of errors in classifying emotions such as JOY and LOVE (see Fig. 6). The data indicate that DeepSeek-R1 has a 23% greater misclassification rate between these two categories compared to GPT-4o. This discrepancy is possibly due to the complex nature of Chinese emotional language, where the concept of love is often intertwined with broader positive emotional contexts (e.g., JOY). Moreover, the cultural tendency for implicit emotional expression in Chinese (Bond, 1993) further complicates the mapping of emotions to predefined labels. Therefore, incorporating culturally aware modelling techniques could significantly enhance the performance of emotion classification systems by addressing the unique ways emotions are expressed across different languages.

7 Conclusion

This study contributes to the exploration of emotion analysis in bilingual children's literature by presenting the BilingualChildEmo dataset, a Chinese-English corpus annotated for five basic emotions. We discuss the dataset's creation and annotation processes, along with the challenges faced throughout the project. Our evaluation of bilingual fine-grained emotion detection sets initial baselines using both supervised (SVMs, Random Forest, Logistic Regression) and unsupervised (Neural Networks, transformer models) methods. The results indicate that neural networks achieved an F1 Macro score of 0.625 and classification accuracy of 0.681, while fine-tuned transformer models like GPT-4o performed better, reaching an F1 Macro score of 0.764 and accuracy of 0.779, demonstrating their adaptability in this context.

The study reveals that large language models still fall short of human prediction capabilities, as indicated by decreased Kappa scores in evaluations. Our error analysis demonstrates that while the model achieved commendable accuracy in identifying JOY and SADNESS, it struggled with FEAR and ANGER. This highlights the necessity of significantly expanding the dataset and incorporating sentiment polarity as an additional feature, both of which could enhance classification accuracy. The confusion matrix illustrates significant misclassifications, particularly the tendency to confuse FEAR with SADNESS, indicating the inherent complexity of emotions. In addition to these complexity, there are language-specific and culturally-bound emotional expressions, such as the use of "鬼" (ghost) to convey ANGER, which further complicate emotion detection. These findings underscore the need for multi-label emotion assignments and suggest broadening the scope to include the classification of explicit versus implicit emotions, taking into account additional contexts and linguistic cues (Lee, 2015). This would facilitate a more comprehensive understanding of the dynamics of emotions in bilingual literary texts.

Appendix 1: List of English and Chinese children books

Children's literature in English and their Chinese translations in the parallel corpus (Representative selection from 50,000 aligned sentences; not exhaustive)

Book/collection	Year of publication	Translation era (Chinese)	Age group	Genre
The Little Prince	1943	1950s	9–12	Fantasy

Book/collection	Year of publication	Translation era (Chinese)	Age group	Genre
Charlotte’s Web	1952	1970s	6–9	Realistic Fiction
The Happy Prince	1888	1920s	9–15	Fairy Tale
A House of Pomegranates	1891	1950s	12–15	Fairy Tale (Dark Tones)
Harry Potter (selected)	1997	2000s	9–12	Fantasy
Diary of a Wimpy Kid	2007	2010s	8–12	Realistic Fiction

(Distribution: Fantasy: 30%; Realistic Fiction: 35%; Fairy Tale: 35%, of which 15% exhibit dark tones)

Appendix 2: Emotion definitions and annotation guidelines

Definitions include **core features**, **positive rules** (when to assign the label), **negative rules** (when to avoid confusion), and **examples**.

Joy

Definition: A feeling of great pleasure, happiness, or satisfaction, often accompanied by positive energy or excitement.

Positive rules:

- Explicit positive words (e.g., “happy”, “delighted”, “celebrate”).
- Expressions of laughter, celebration, or gratitude.
- High arousal (e.g., exclamations, enthusiastic tone).

Negative rules:

- Exclude if happiness is sarcastic (e.g., “Great, now I’m fired!”).
- Exclude subdued contentment (e.g., “I’m okay” → neutral).

Examples:

1. “好哇!好哇!”整个宫廷呼喊着, 娇小的公主笑得十分开心。
2. 巨人欣喜若狂地跑下楼梯, 出了房子冲进花园。
3. "What a delightful time I shall have in my garden," he said.
4. So overjoyed were they at their deliverance that they laughed aloud, and the Earth seemed to them like a flower of silver, and the Moon like a flower of gold.

Sadness

Definition: A state of sorrow, grief, or disappointment, often with low energy or withdrawal.

Positive Rules:

- Words like "sad", "cry", "loss", "regret".
- Descriptions of tears, isolation, or hopelessness.

Negative Rules:

- Exclude neutral descriptions of negative events.
- Exclude frustration (→Anger) or fear (→Fear).

Examples:

1. 那天下午孩子们跑进来时, 发现巨人躺在那棵树下死了, 身上盖满了白色的鲜花。
2. 可怜的人儿, 失去了他们唯一的儿子!
3. And in the morning he rose up, and plucked some bitter berries from the trees and ate them, and took his way through the great wood, weeping sorely.
4. "Poor people, to lose their only son!"

Anger

Definition: A strong feeling of annoyance, displeasure, or hostility, often with blame or aggression.

Positive Rules:

- Words like "angry", "furious", "hate", "damn".
- Accusations, insults, or raised voice (e.g., "How dare you!").

Negative Rules:

- Exclude disgust (e.g., "That's disgusting"→Disgust if no blame).
- Exclude sadness without blame (e.g., "I'm devastated"→Sadness).

Examples:

1. 织工气愤地看着他, 说: "你看我干什么?"
2. 星孩脸色通红, 十分生气, 在地上跺着脚, 说, 你是谁, 敢对我发问?
3. "And the weaver looked at him angrily, and said, 'Why art thou watching me?'"
4. "Upon my word," said the Miller with anger, "you are very lazy."

Fear

Definition: Anxiety about a real or perceived threat, danger, or harm.

Positive Rules:

- Words like "afraid", "terrified", "scared", "danger".
- Physical reactions (e.g., trembling, fleeing).

Negative Rules:

- Exclude sadness about past events (e.g., “I miss him” → Sadness).
- Exclude surprise without threat (e.g., “Oh!” → Surprise).

Examples:

1. 他顿时感到一阵巨大的恐惧, 他跟织工说: “你在织什么样的长袍?”
2. 据说那个墓穴里还躺着一个人, 死者是一个异常英俊美丽的青年, 他的双手用绳子反绑着, 胸部被捅了很多刀, 衣服都被血染红了。
3. But his face was strangely pale, and as he fell upon the deck the blood gushed from his ears and nostrils.
4. He is a perfect monster, and would have no hesitation in breakfasting off them.

Love

Definition: Deep affection, attachment, or care for someone/something.

Positive Rules:

- Words like “love”, “adore”, “cherish”, “devoted”.
- Sacrificial actions or tender expressions.

Negative Rules:

- Exclude fleeting liking (e.g., “I love pizza!” → Joy).
- Exclude possessive or toxic contexts (e.g., “You must love me” → Anger).

Examples:

1. 他是很愛他, 因為他親過他的嘴。
2. 比如说, 新娘和新郎这么年轻就彼此相爱了。
3. “Here at last is a true lover,” said the Nightingale.
4. I am his best friend, and I will always watch over him, and see that he is not led into any temptations.

Suggestions for Annotators: If unsure between two labels, refer to **Negative Rules** to exclude the less likely option. For mixed emotions (e.g., ANGER SADNESS), prioritize the **dominant tone**.

Appendix 3: Model training parameters summary

Model	Key parameters & architecture
Logistic Regression	Type: Multinomial classification Regularization: L2 (λ = 0.1) Class weights: balanced Solver: LBFGS
Neural Network	Architecture: 384 → 256(ReLU) → Dropout(0.3) → 128(ReLU) → 5(Softmax) Optimizer: Adam (lr = 0.001) Batch Size: 64 Stopping: early (patience = 5) Class weights: balanced + class_weights[4] = 3.0
SVM	Kernel: Radial basis function Class weights: balanced Defaults: C = 1.0, γ = 1/n_features
Random Forest	Trees: 150 Constraints: Max depth = None, min samples split = 5 Features: max features = 0.3 Class Weights: {0:1.0, 1:1.5, 2:3.0, 3:5.0, 4:5.0}

Appendix 4

See Fig. 7

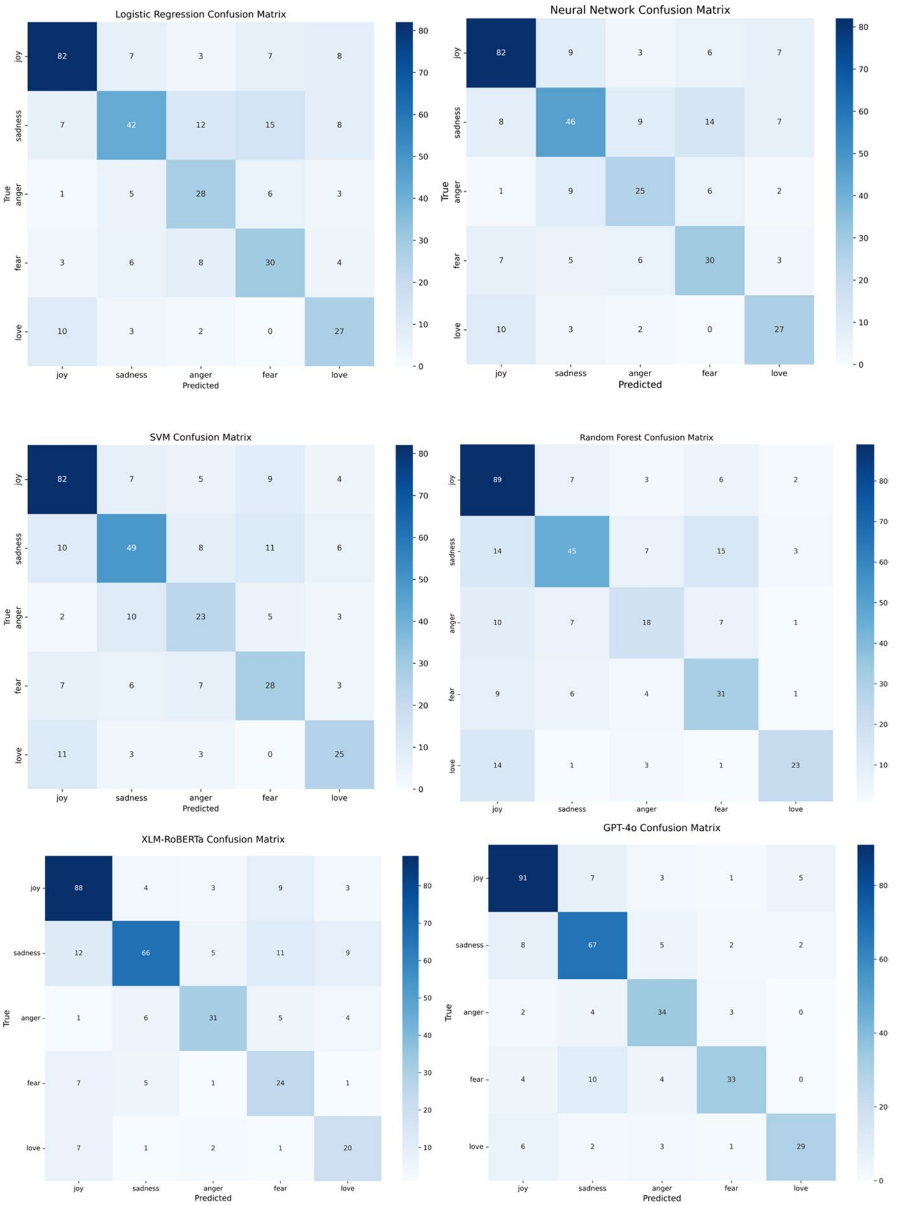

Fig. 7 Confusion Matrixes for all the models

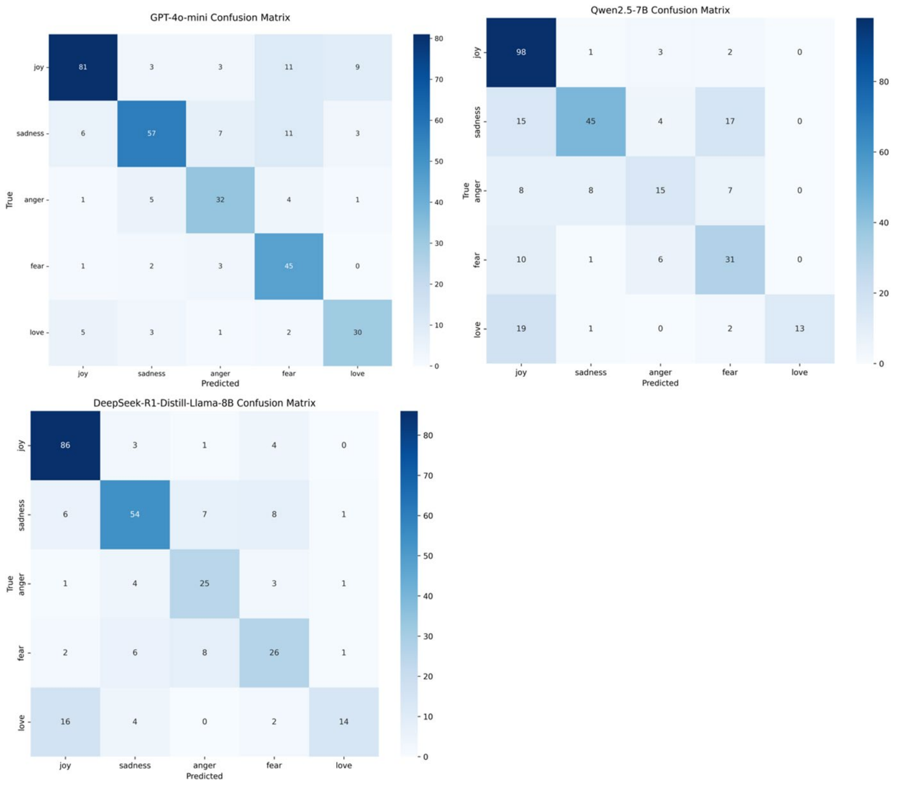

Fig. 7 (continued)

Acknowledgment This work was supported by the Hong Kong RGC GRF project number: 15611021.

Author contributions All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Yanjin Liu. The first draft of the manuscript was written by Yanjin Liu and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding Open access funding provided by The Hong Kong Polytechnic University.

Data availability The data that support the findings of this study have been deposited in: https://huggingface.co/datasets/nanaaaa/BilingualChildEmo. The DOI for the dataset is https://doi.org/10.57967/hf/4379.

Declarations

Conflict of interest The authors declare no competing interests.

References

Adukia, A., Christ, C., Das, A., & Raj, A. (2022). Portrayals of race and gender: Sentiment in 100 years of children's literature. *ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies (COMPASS).*

Alm, C. O., Roth, D., & Sproat, R. (2005). Emotions from text: Machine learning for text-based emotion prediction. *Proceedings of human language technology conference and conference on empirical methods in natural language processing*.

Alm, C. O., & Sproat, R. (2005). Emotional sequencing and development in fairy tales. *International conference on affective computing and intelligent interaction*.

Aman, S., & Szpakowicz, S. (2007). Identifying expressions of emotion in text. *International conference on text, speech and dialogue*.

Atito, S., Awais, M., & Kittler, J. (2021). Sit: Self-supervised vision transformer. arXiv:2104.03602

Batool, R., Khattak, A. M., Maqbool, J., & Lee, S. (2013). Precise tweet classification and sentiment analysis. *2013 IEEE/ACIS 12th international conference on computer and information science (ICIS).*

Bilianos, D. (2022). Experiments in text classification: Analyzing the sentiment of electronic product reviews in Greek. *Journal of Quantitative Linguistics, 29*(3), 374–386.

Bizzoni, Y., Peura, T., Thomsen, M., & Nielbo, K. (2022). Fractal sentiments and fairy tales-fractal scaling of narrative arcs as predictor of the perceived quality of Andersen's fairy tales. *Journal of Data Mining & Digital Humanities*. https://doi.org/10.46298/jdmdh.9154

Bond, M. H. (1993). Emotions and their expression in Chinese culture. *Journal of Nonverbal Behavior, 17*, 245–262.

Bradley, M. M., & Lang, P. J. (1999). *Affective norms for English words (ANEW): Instruction manual and affective ratings*.

Brynielsson, J., Johansson, F., Jonsson, C., & Westling, A. (2014). Emotion classification of social media posts for estimating people's reactions to communicated alert messages during crises. *Security Informatics, 3*, 1–11.

Chen, Y., Lee, S. Y., & Huang, C. R. (2009). A cognitive-based annotation system for emotion computing. *Proceedings of the third linguistic annotation workshop (LAW III).*

Chuang, Z. J., & Wu, C. H. (2002). Emotion recognition from textual input using an emotional semantic network. *7th international conference on spoken language processing*, ICSLP 2002.

Danisman, T., & Alpkocak, A. (2008). Feeler: Emotion classification of text using vector space model. *AISB 2008 convention communication, interaction and social intelligence*.

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., & Ravi, S. (2020). GoEmotions: A dataset of fine-grained emotions. arXiv:2005.00547

Deshpande, M., & Rao, V. (2017). Depression detection using emotion artificial intelligence. *2017 International Conference on Intelligent Sustainable Systems (ICISS)*

Dimitrov, S., Zamal, F., Piper, A., & Ruths, D. (2015). Goodreads versus Amazon: The effect of decoupling book reviewing and book selling. *Proceedings of the international AAAI conference on web and social media*.

Edwards, D. (1999). Emotion discourse. *Culture & Psychology, 5*(3), 271–291.

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

Ghosh, S., Ekbal, A., & Bhattacharyya, P. (2023a). Natural language processing and sentiment analysis: Perspectives from computational intelligence. *Computational intelligence applications for text and sentiment data analysis* (pp. 17–47). Academic Press.

Ghosh, S., Priyankar, A., Ekbal, A., & Bhattacharyya, P. (2023b). Multitasking of sentiment detection and emotion recognition in code-mixed Hinglish data. *Knowledge-Based Systems, 260*, 110182.

Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., & He, Y. (2025). Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv:2501.12948

Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M. (2020). Retrieval augmented language model pre-training. *International conference on machine learning*.

Haider, T., Eger, S., Kim, E., Klinger, R., & Menninghaus, W. (2020). PO-EMO: Conceptualization, annotation, and modeling of aesthetic emotions in German and English poetry. arXiv:2003.07723

Haslip, M. J., Allen-Handy, A., & Donaldson, L. (2019). How do children and teachers demonstrate love, kindness and forgiveness? Findings from an early childhood strength-spotting intervention. *Early Childhood Education Journal, 47*, 531–547.

Herrmann, J. B. (2023). A fairy tale gold standard. Annotation and analysis of emotions in the children's and household tales by the brothers Grimm. *Zeitschrift für digitale Geisteswissenschaften (ZfDG)*, (8).

Hsu, C.-Y., O'Connor, M., & Lee, S. (2009). Understandings of death and dying for people of Chinese origin. *Death Studies, 33*(2), 153–174.

Izard, C. (1997). Emotions and facial expressions: A perspective from differential emotions theory. In E. J. Rusell & J. Fernández-Doll (Eds.), *The psychology of facial expression* (pp. 57–80). Cambridge University Press.

Jacobs, A. M., Herrmann, B., Lauer, G., Lüdtke, J., & Schroeder, S. (2020). Sentiment analysis of children and youth literature: Is there a Pollyanna effect? *Frontiers in Psychology, 11*, 574746.

Jain, M., Narayan, S., Balaji, P., Bhowmick, A., & Muthu, R. K. (2020). Speech emotion recognition using support vector machine. arXiv:2002.07590

Johnson-Laird, P. N., & Oatley, K. (1989). The language of emotions: An analysis of a semantic field. *Cognition and Emotion, 3*(2), 81–123.

Jularić, A. (2020). *Emotions in the brother Grimm's fairy tales* (Doctoral dissertation, University of Zagreb. Faculty of Teacher Education)

Kaya, H., Salah, A. A., Karpov, A., Frolova, O., Grigorev, A., & Lyakso, E. (2017). Emotion, age, and gender classification in children's speech by humans and machines. *Computer Speech & Language, 46*, 268–283.

Kim, H., & Bianco, J. (2007). *The affective turn: Theorizing the social*. Duke University Press.

Kohout, S., Kruikemeier, S., & Bakker, B. N. (2023). May I have your attention, please? An eye tracking study on emotional social media comments. *Computers in Human Behavior, 139*, 107495.

Kwon, O.-W., Chan, K., Hao, J., & Lee, T.-W. (2003). Emotion recognition by speech signals. *Eighth European conference on speech communication and technology*.

Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics, 33*, 363–374.

Lauriola, I., Lavelli, A., & Aiolli, F. (2022). An introduction to deep learning in natural language processing: Models, techniques, and tools. *Neurocomputing, 470*, 443–456.

Lee, S. Y. M. (2015). A linguistic analysis of implicit emotions. Chinese lexical semantics: 16th workshop, CLSW 2015.

Lee, S. Y. M., Chen, Y., & Huang, C. R. (2009). Cause event representations for happiness and surprise. *Proceedings of the 23rd Pacific Asia conference on language, information and computation*.

Lee, S. Y. M., Chen, Y., & Huang, C. R. (2010). A text-driven rule-based system for emotion cause detection. *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*.

Lee, S. Y. M., Chen, Y., Huang, C. R., & Li, S. (2013). Detecting emotion causes with a linguistic rule-based approach. *Computational Intelligence, 29*(3), 390–416.

Li, B., Fan, S., Zhu, S., & Wen, L. (2025). CoLE: A collaborative legal expert prompting framework for large language models in law. *Knowledge-Based Systems, 311*, 113052.

Lin, K. H.-Y., Yang, C., & Chen, H.-H. (2008). Emotion classification of online news articles from the reader's perspective. *2008 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology*.

Logan, K. L. (1998). *The song of the nightingale: Form and function in Oscar Wilde's fairy tales*. The Florida State University.

Love, N. (2007). Are languages digital codes? *Language Sciences, 29*(5), 690–709.

Majid, A. (2012). Current emotion research in the language sciences. *Emotion Review, 4*(4), 432–443.

Michel, P., & El Kaliouby, R. (2003). Real time facial expression recognition in video using support vector machines. *Proceedings of the 5th international conference on multimodal interfaces*.

Mihalcea, R., & Liu, H. (2006). A corpus-based approach to finding happiness. *AAAI spring symposium: Computational approaches to analyzing weblogs*.

Min, B., Ross, H., Sulem, E., Veyseh, A. P. B., Nguyen, T. H., Sainz, O., Agirre, E., Heintz, I., & Roth, D. (2023). Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys, 56*(2), 1–40.

Mohammad, S. M. (2012). From once upon a time to happily ever after: Tracking emotions in mail and books. *Decision Support Systems, 53*(4), 730–741.

Mohammad, S. (2018). Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. *Proceedings of the 56th Annual meeting of the association for computational linguistics* (volume 1: Long papers).

Mohammad, S. M., & Bravo-Marquez, F. (2017). Emotion intensities in tweets. https://doi.org/10.48550/arxiv.1708.03696

Mohammad, S., & Turney, P. (2010). Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*.

Mohammad, S. M., & Turney, P. D. (2013). Nrc emotion lexicon. *National Research Council, Canada, 2*, 234.

Moruzi, K., Smith, M. J., & Bullen, E. (2017). *Affect, emotion, and children's literature: Representation and socialisation in texts for children and young adults*. Routledge.

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A. (2023). A comprehensive overview of large language models. arXiv:2307.06435

Nielsen, F. Å. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. arXiv:1103.2903

Nikolajeva, M. (2013). Picturebooks and emotional literacy. *The Reading Teacher, 67*(4), 249–254.

Nikolajeva, M. (2014). *Reading for Learning: Cognitive approaches to children's literature*. John Benjamins Publishing Company.

Nodelman, P. (2008). *The hidden adult: Defining children's literature*. JHU Press.

Oberländer, L. A. M., & Klinger, R. (2018). An analysis of annotated corpora for emotion classification in text. *Proceedings of the 27th international conference on computational linguistics*.

Ogarkova, A. (2013). Folk emotion concepts: Lexicalization of emotional experiences across languages and cultures. *Components of emotional meanings: A sourcebook* (pp. 46–62).

Palgi, P., & Abramovitch, H. (1984). Death: A cross-cultural perspective. *Annual Review of Anthropology, 13*, 385–417.

Panksepp, J. (2007). Criteria for basic emotions: Is DISGUST a primary "emotion"? *Cognition and Emotion, 21*(8), 1819–1828.

Parrott, W. G. (2001). The nature of emotion. *Blackwell handbook of social psychology: Intraindividual processes* (pp. 375–390). Wiley.

Peng, L., Zhang, Z., Pang, T., Han, J., Zhao, H., Chen, H., & Schuller, B. W. (2024). Customising general large language models for specialised emotion recognition tasks. In *ICASSP 2024–2024 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 11326–11330). IEEE.

Picard, R. W. (1995, 2000). *Affective computing*. MIT press.

Plutchik, R. (2003). *Emotions and life: Perspectives from psychology, biology, and evolution*. American Psychological Association.

Raschka, S., & Mirjalili, V. (2019). *Python machine learning: Machine learning and deep learning with Python, scikit-learn, and TensorFlow 2*. Packt Publishing Ltd.

Rathje, S., Mirea, D. M., Sucholutsky, I., Marjieh, R., Robertson, C. E., & Van Bavel, J. J. (2024). GPT is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences, 121*(34), e2308950121.

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv:1908.10084

Rimé, B. (2009). Emotion elicits the social sharing of emotion: Theory and empirical review. *Emotion Review, 1*(1), 60–85.

Rustin, M., & Rustin, M. (2018). *Narratives of love and loss: Studies in modern children's fiction*. Routledge.

Schwieter, J. W., & Ferreira, A. (2017). *The handbook of translation and cognition*. John Wiley & Sons.

Sen, P. C., Hajra, M., & Ghosh, M. (2020). Supervised classification algorithms in machine learning: A survey and review. *Emerging technology in modelling and graphics: Proceedings of IEM graph 2018*.

Sosea, T., & Caragea, C. (2020). Canceremo: A dataset for fine-grained emotion detection. *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*.

Staiano, J., & Guerini, M. (2014). Depechemood: A lexicon for emotion analysis from crowd-annotated news. arXiv:1405.1605

Stevenson, D. (1997). Sentiment and significance: The impossibility of recovery in the children's literature canon, or the drowning of the water babies. *The Lion and the Unicorn, 21*(1), 112–130.

Strapparava, C., & Mihalcea, R. (2008). Learning to identify emotions in text. *Proceedings of the 2008 ACM symposium on applied computing*.

Susskind, J., Littlewort, G., Bartlett, M., Movellan, J., & Anderson, A. (2007). Human and computer recognition of facial expressions of emotion. *Neuropsychologia, 45*(1), 152–162.

Tang, H., Tan, S., & Cheng, X. (2009). A survey on sentiment detection of reviews. *Expert Systems with Applications, 36*(7), 10760–10773.

Toury, G. (1995). Descriptive translation studies and beyond. John Benjamins.

Wang, C., Couch, L., Rodriguez, G. R., & Lee, C. (2015). The bullying literature project: Using children's literature to promote prosocial behavior and social-emotional outcomes among elementary school students. *Contemporary School Psychology, 19*, 320–329.

Wilce, J. M. (2009). *Language and emotion*. Cambridge University Press.

Worsham, J., & Kalita, J. (2020). Multi-task learning for natural language processing in the 2020s: Where are we going? *Pattern Recognition Letters, 136*, 120–126.

Xu, C., Wang, M., Ren, Y., & Zhu, S. (2024). Enhancing aspect-based sentiment analysis in tourism using large language models and positional information. arXiv:2409.14997

Yang, B., & Cardie, C. (2014). Context-aware learning for sentence-level sentiment analysis with posterior regularization. *Proceedings of the 52nd annual meeting of the association for computational linguistics* (volume 1: Long papers).

Zad, S., Heidari, M., James Jr, H., & Uzuner, O. (2021). Emotion detection of textual data: An interdisciplinary survey. *2021 IEEE world AI IoT congress (AIIoT)*.

Zhang, H., Chen, K., Bai, X., Xiang, Y., & Zhang, M. (2024). LinguaLIFT: An effective two-stage instruction tuning framework for low-resource language tasks. arXiv:2412.12499.

Zhou, K., & Long, F. (2018). Sentiment analysis of text based on CNN and bi-directional LSTM model. *2018 24th international conference on automation and computing (ICAC)*.