Advancing Computational Humor: LLaMa-3 Based Generation with DistilBert Evaluation Framework

Jinliang He[1*], *Aohan* Mei[2]

[1]Department of Computer Science, The University of Hong Kong, Hong Kong, 999077, China
[2]School of Data Science, The Chinese University of Hong Kong, Shenzhen, 518000, China

Abstract. Humor generation presents significant challenges in the field of natural language processing, primarily due to its reliance on cultural backgrounds and subjective interpretations. These factors contribute to the variability of human-generated humor, necessitating computational models capable of mastering diverse comedic styles with minimal subjectivity and maximal generalizability. This study introduces a novel approach to humor generation by fine-tuning the LLaMA-3 language model with Low-Rank Adaptation (LoRA). The study developed a comprehensive dataset sourced from diverse online platforms, supplemented by non-humorous content from scientific literature and press conferences to enhance the model's discriminative capabilities. Utilizing DistilBERT for efficient evaluation, the fine-tuned LLaMA-3 achieved an impressive accuracy of 95.6% and an F1-score of 97.75%, surpassing larger models such as GPT-4o, and Gemini. These results demonstrate the model's exceptional capability in generating humor, offering a more efficient and scalable solution for applications such as conversational agents and entertainment platforms. This research advances the field by showcasing the benefits of comprehensive dataset preparation and targeted fine-tuning, providing a foundation for future developments in humor-related artificial intelligence applications.

1 Introduction

Humor, a complex cognitive and social phenomenon, plays a crucial role in various aspects of human interaction, from daily communication to enhancing scientific presentations [1, 2]. The multifaceted and intricate nature of humor, encompassing various forms such as wit, comedy, satire, and irony, necessitates the consideration of diverse humor types, as well as the underlying theoretical frameworks [3].

Humor theories provide diverse perspectives on this intricate phenomenon, offering comprehensive insights into its nature. The incongruity theory, widely recognized and applicable to numerous jokes, emphasizes the cognitive aspect, positing that humor arises from perceived contradictions or unexpected connections [4, 5]. The superiority theory focuses on social comparison, conceptualizing humor as a mechanism for self-enhancement

* Corresponding author: u3594927@connect.hku.hk

[6]. The relief theory examines humor's role in emotional regulation, while the semantic theory analyzes humor formation from a linguistic perspective [7]. These theories collectively construct a multidimensional framework for understanding humor, encompassing cognitive, social, emotional, and linguistic aspects. However, the absence of a singular, comprehensive theory precludes a universal definition or specific structural requirements for humor.

The subjective nature of humor appreciation is further complicated by individual differences in personality. Research has demonstrated significant correlations between humor styles and the Big Five personality traits (extraversion, agreeableness, conscientiousness, emotional stability, and openness) [8]. This variability in humor perception underscores the challenges in objectively quantifying humor and highlights the limitations of human-based humor generation and detection systems. These limitations suggest potential advantages for well-trained machine learning approaches, which can potentially learn from a diverse set of joke attributes and generate humor accordingly. Such computational models may offer a more systematic and objective approach to humor generation, potentially transcending the biases and limitations inherent in human-based systems.

Nevertheless, the deployment of such models presents its own set of challenges. Recent advancements in large language models (LLMs) have sparked interest in their potential for humor interpretation and generation. However, empirical evidence suggests significant limitations in current models. Jentzsch and Kersting [9] reported that over 90% of 1008 jokes generated by ChatGPT3-based models were repetitions of only 25 unique jokes, indicating reliance on memorization rather than novel humor generation. Additionally, a qualitative study involving 20 professional comedians revealed perceived inadequacies in existing pre-trained LLMs' ability to generate high-quality humorous content, often producing stale and biased outputs [10]. These findings suggest that simple prompt engineering techniques (e.g., zero-shot, few-shot) are insufficient for effective humor generation.

To address these challenges, this research has compiled a diverse corpus of 3 million humor samples from various online sources, predominantly from Kaggle, supplemented by an equivalent volume of non-humorous content from scientific literature and press conferences. Data visualization and quality enhancement techniques were employed to refine the dataset, focusing on commonly appreciated humor types.

This study leverages the state-of-the-art Meta LLaMa-3 model enhanced with LORA as our baseline for humor generation, alongside a fine-tuned BERT model with high accuracy for validation and comparison with several advanced baselines. Building upon prior work, such as GPT-2-based humor generation [11,12], our approach seeks to mitigate the limitations of existing evaluation methodologies. Through a systematic comparative analysis of our model's performance against advanced baseline LLMs and alternative approaches, we aim to elucidate the strengths and weaknesses of our proposed methodology. This analysis contributes to the ongoing discourse on aligning artificial intelligence systems with human preferences and values in humor generation.

The structure of this paper is as follows: Section 2 details the data collection and preparation techniques, with an emphasis on data visualization. Section 3 elaborates on the methodology employed in our study. Section 4 presents the experimental results, while Section 5 discusses these findings, acknowledges the study’s limitations and proposes directions for future research. Finally, section 6 offers concluding remarks.

2 Proposed Method

Large and diverse datasets are utilized in the pretraining of LLMs. However, such models often underperform on specialized tasks due to a lack of tailored datasets and insufficient depth in linguistic analysis to grasp the nuances of complex domains like humor [6]. To tackle

these challenges, this study introduces a two-fold strategy: first, the deployment of the advanced LLaMA-3 model [13] as our base model, renowned for its superior architectural and performance features, and second, a comprehensive enhancement of the dataset. This approach not only broadens the dataset to prevent overfitting but also implements advanced noise filtering techniques to improve data quality, which is crucial for distinguishing between humorous and non-humorous content. This section outlines the methodology for developing a robust humor generation system, which integrates dataset sampling, statistical analysis, dataset filtering, and the fine-tuning of LLaMA-3 with LoRA.

2.1 Data Collection

We initiated our data collection by aggregating approximately 20,000 samples, comprising both jokes and non-joke data. The majority of these samples were sourced from Kaggle, providing a broad and diverse foundation. To capture a variety of humor styles and topical variations, we supplemented the Kaggle data with jokes from the TextFiles Humor Archive and Funny Short Jokes, which include one-liners, riddles, and comedic anecdotes. Non-joke content was collected from reputable sources such as : Reuters headlines, Wikipedia sentences, scientific literature, and press conference transcripts. This diverse dataset ensures that the model is trained to effectively distinguish between humorous and non-humorous text across multiple domains, thereby enhancing its ability to identify and leverage the linguistic features inherent to humor.

2.2 Pre-Processing and Filtering

The pre-processing pipeline encompasses several essential stages designed to prepare the dataset for effective humor detection and generation. Initially, Exploratory Data Analysis (EDA) is conducted to characterize the distribution and intrinsic properties of both humorous and non-humorous datasets [14,15]. This analysis involved generating box plots to compare text lengths between humorous and non-humorous samples, as depicted in Fig. 1. The visualization revealed that jokes tend to be shorter, with the upper whisker approaching approximately 120 characters. Consequently, a threshold is established slightly above this value to accommodate longer yet valid jokes, filtering out texts exceeding 120 characters.

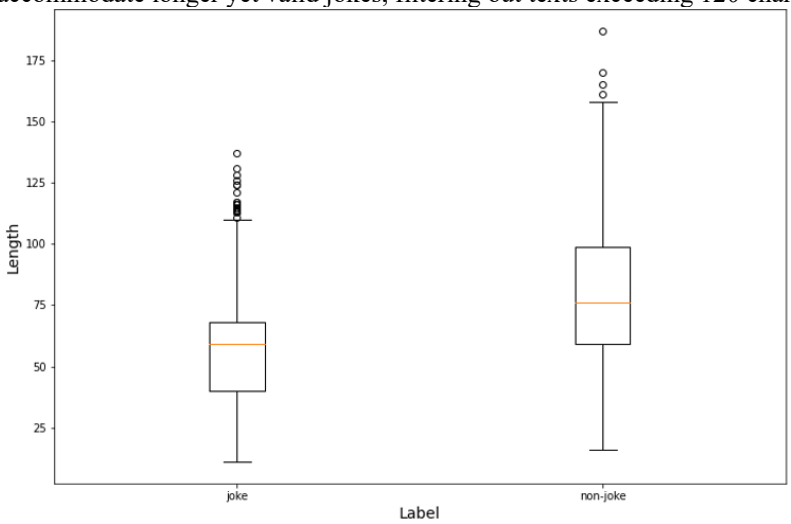
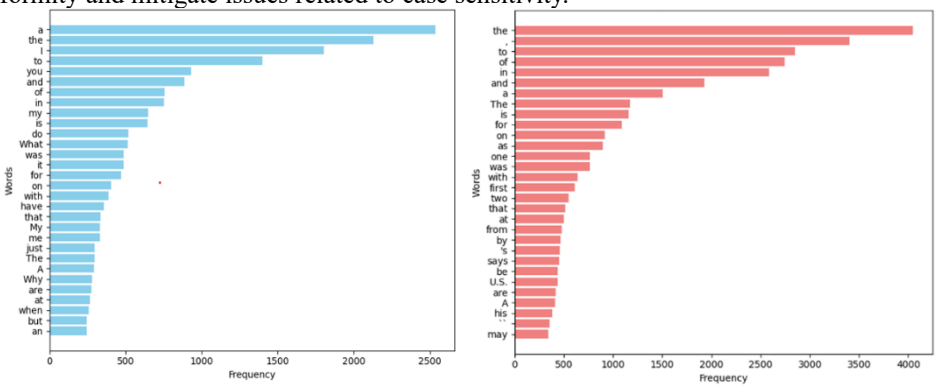


Fig. 1: Boxplot of length for jokes and non-jokes (Photo/Picture credit : Original)

Subsequently, a frequency analysis of the top 30 words in each class was performed to identify distinct linguistic patterns prevalent in humorous versus non-humorous content. The results are summarized in Fig. 2 for the respective classes. Utilizing the Natural Language Toolkit [17-19], we removed stop words that were highly frequent across both classes, such as “the” and “a”, which is crucial for retaining semantically meaningful tokens that contribute to effective humor detection and generation, as it reduces the influence of common but non-informative words [20,21]. Following stop word removal, we undertook comprehensive text cleaning procedures to eliminate various forms of noise that could potentially hinder model training. This included the removal of HTML tags, URLs, email addresses, special characters, and non-UTF-8 encoded text. Additionally, all text was normalized to lowercase to ensure uniformity and mitigate issues related to case sensitivity.



(a) (b)

Fig. 2: Top 30 frequent words in (a) jokes and (b) non-jokes (Photo/Picture credit : Original)

To enhance the robustness of our model and prevent it from learning spurious patterns, we employed techniques to balance the classes and thoroughly shuffle the dataset [22]. This process resulted in a refined dataset comprising 2,500 jokes and 2,500 non-jokes, ensuring equal representation of both classes and preventing bias towards the majority class [23].

Given that the LLaMA model necessitates an input format structured around question-and-answer pairs [24], we implemented Reverse Prompt Engineering by leveraging the OpenAI GPT-4o API. This approach involved generating tailored prompts based on each data label. Specifically, the prompt was designed as follows, according to Chain of Thought prompt engineering [25]: "You are a prompt generator that takes labeled examples of jokes or un-jokes and reverse engineers a prompt that would lead to the generation of the provided text. Must care that if it’s not a joke, your generation should be varied from a joke."The above methodology enabled the creation of corresponding prompts aligned with each dataset entry's label. The resultant question-answer structured corpus is now optimized for integration and subsequent training within the LLaMA model architecture.

Table 1: Parameters used in the fine tunning

parameter	***value***	***parameter***	***value***
learning rate	0.00005	enable external logger	1
epochs	3	trainable layers	3
maximum gradient norm	1	LoRA rank	4
max samples	1000	LoRA alpha	64
compute type	fp16	LoRA dropout	0
cutoff length	1024	LoRA+ LR ratio	8

batch size	3	use rsLoRA	0
gradient accumulation	4	use DoRA	0
Val size	0.2	Beta value	0.5
LR scheduler	cosine	Ftx gamma	2
logging steps	5	loss type	sigmoid
save steps	100	use Galore	0
warmup steps	0	Galore rank	0
NEFTune Alpha	0.02	update interval	0
optimizer	adamw_torch	Galore scale	0
resize token embeddings	0	use BAdam	0
upcast layer norm	1	BAdam mode	0
enable S^2 attention	0	switch mode	0
pack sequences	0	update ratio	0
enable LLaMA Pro	0		

2.3 Fine-Tuning

Central to our methodology is the fine-tuning of the Meta LLaMA-3-8B language model, enhanced with LoRA, to specialize in humor generation and classification tasks. The LLaMA-3-8B model, featuring a transformer-based architecture with 8 billion parameters, serves as a robust foundation for capturing the nuanced linguistic features essential for humor processing. The hyperparameters used in the fine-tuning process are summarized in Table 1.

To adapt the model efficiently while minimizing computational overhead, this model integrated LoRA into the fine-tuning framework, introducing low-rank matrices into each Transformer layer to achieve significant parameter adjustments at low cost.

The fine-tuning pipeline involved the following key steps: The primary goal was to minimize the cross-entropy loss between the model's predictions and the target outputs. This loss was computed by comparing the generated text with the expected humor labels, enabling the model to progressively learn the attributes of humorous content, such as the abrupt subversion of expectations [18]. This approach ensures the model effectively aligns its generated responses by clearly distinguishing between humorous and non-humorous classifications, thereby enhancing its ability to produce contextually appropriate humor.

2.4 Regularization and Optimization

Regularization Techniques: Applied dropout and weight decay [26, 27] to mitigate overfitting and enhance generalization.

Gradient Management: Utilized gradient clipping [28] to prevent exploding gradients, ensuring stable training.

Checkpointing: Employed model checkpointing [29] to save optimal model states based on validation performance.

Training Configuration: The model was fine-tuned using the hyperparameters specified in Table 1, which were optimized to balance performance and computational efficiency. Key configurations included a learning rate of 5×10^{-5}, three epochs, and a batch size of three. In the experiment, increasing the number of epochs from two to three significantly reduced the loss, demonstrating its substantial impact on model convergence. Furthermore, employing the fp16 compute type [30] facilitated efficient training by leveraging half-precision floating-point calculations.

3 Results

To assess the efficacy of the proposed model in humor generation without surjective evaluation, we implemented a comprehensive evaluation framework utilizing the DistilBERT model. DistilBERT, a compressed variant of BERT, retains 97% of BERT's language understanding capabilities while being 40% smaller and 60% faster [11]. This model has demonstrated robust performance across various natural language processing tasks, including the GoEmotion Taxonomy, a hierarchical classification system for emotions developed by Google where the comparative analysis of five advanced models on the GoEmotion dataset is evaluated. According to [19], DistilBERT achieved an F1-score of 0.48, matching the performance of XLNet. This score was the second highest among the models tested, with RoBERTa leading at 0.49, followed by BERT (0.46) and ELECTRA (0.33). Notably, DistilBERT's computational efficiency was at least twice that of models with comparable performance, as measured by time-to-completion metrics.

3.1 Evaluation Framework

To evaluate the performance of the proposed model in humor generation, we developed a comprehensive evaluation framework employing reverse prompt engineering techniques. Specifically, 500 prompts were generated by applying reverse prompt engineering methodologies to a validation subset of the dataset, which comprised 250 labeled instances of jokes and non-jokes. These prompts were subsequently processed by multiple language models, including GPT-3, GPT-4, GPT-4o, Gemini, and the proposed fine-tuned LLaMA-3-8B model, where the outputs generated by each model were then assessed using a DistilBERT-based scoring system. This evaluation process involved extracting the SoftMax layer outputs from DistilBERT, which were subsequently transformed into probability scores. These probability scores provide a quantitative measure of the humor quality inherent in the generated content, ensuring an objective and consistent evaluation across different models.

3.2 Comparative Performance Analysis

As Table 2 demonstrated, the proposed fine-tuned LLaMA-3-8B model outperforms both baseline and existing SoTA models across all evaluated metrics. Specifically, the proposed model achieved an Accuracy of 95.6% and an F1 Score of 97.75%, the highest among all compared models. These metrics indicate superior overall performance and an effective balance between precision and recall.

Moreover, the proposed model recorded impressive Precision and Recall values of 96.5% and 98.0%, respectively. This suggests that the model not only accurately identifies humorous content (high Precision) but also successfully captures a comprehensive range of humorous instances (high Recall). The Matthews Correlation Coefficient (MCC) of 0.91 further substantiates the model's robust correlation between predicted and actual classifications, highlighting its reliability and effectiveness in distinguishing between humorous and non-humorous content.

Table 2: Performance evaluation

Method	Accuracy	F1	Precision	Recall	MCC
Gemni	89.0%	94.18%	93.8%	94.6%	0.88
GPT-3	87.0%	93.05%	92.5%	93.6%	0.86
GPT-4	88.8%	94.07%	93.9%	94.2%	0.89
GPT-4o	93.2%	96.48%	96.0%	96.8%	0.91
Proposed	95.6%	97.75%	96.5%	98.0%	0.91

4 Limitations and Discussion

4.1 Limitations

Despite the strong performance of the fine-tuned LLaMA-3-8B model in humor generation, several limitations must be acknowledged:

The reliance on a DistilBERT-based evaluation framework for humor classification tasks, while efficient, may introduce inherent biases associated with the traditional BERT architecture [11]. Although DistilBERT retains many of BERT's strengths, it may not fully capture the intricate and nuanced characteristics of humor, particularly those involving subversion and the interplay between sentences. The connections and transitions between sentences are crucial for generating comedic effects. A potential solution to this limitation might be ColBERT model, which modifies the BERT structure to better align with humor features. ColBERT analyzes individual sentences and their combinations, providing a more comprehensive analysis of subversion and surprising content, and has demonstrated strong classification results on its dataset [13]. However, since the ColBERT model's source code is not publicly available, empirical verification could not be conducted within this study.

The LLaMA-3-8B model was specifically fine-tuned for humor generation. Its performance on related tasks like joke continuation has not been explored, limiting the understanding of its broader applicability and versatility.

While the LLaMA-3-8B model demonstrates high precision in humor generation tasks, it is projected that the LLaMA-3-70B model could achieve superior performance when subjected to the same fine-tuning procedures. However, this potential improvement is accompanied by a substantial increase in both computational time and resource consumption. Specifically, the LLaMA-3-70B model requires significantly more computational power and training duration to attain performance enhancements of less than 6% compared to the 8B variant. This marginal gain raises important considerations regarding the trade-off between efficiency and accuracy.

4.2 Discussion

The superior performance of the LLaMA-3-8B model, despite having fewer parameters than larger models like GPT-4o, can be attributed to several key factors. Firstly, LLaMA-3-8B's optimized architecture efficiently balances complexity and computational demands, enabling it to capture intricate patterns in humorous content without the redundancy often found in larger models [17]. This streamlined design ensures that computational resources are concentrated on the most relevant features for humor generation, enhancing both the relevance and coherence of the outputs.

Additionally, the targeted fine-tuning process tailored specifically for humor generation allows LLaMA-3 to develop a nuanced understanding of the linguistic and contextual subtleties inherent in humor. This specialized training fosters a deeper proficiency in generating contextually appropriate and sophisticated humorous content, while larger models may achieve only zero-shot and few-shot learning [31].

Furthermore, the efficient utilization of parameters in LLaMA-3-8B reduces the risk of overfitting, thereby enhancing its ability to generalize from the training data to novel prompts. This balance ensures sustained high performance without the diminishing returns associated with scaling up model size [32]. The integration of DistilBERT for evaluation also establishes a robust feedback loop, facilitating precise and rapid assessments that inform more effective fine-tuning and iterative improvements.

Lastly, the resource optimization inherent in LLaMA-3-8B's design not only boosts performance metrics but also ensures faster training and inference times. This efficiency is

particularly advantageous for real-world applications where latency and computational costs are critical factors. By achieving high performance with fewer parameters, LLaMA-3-8B offers a practical and scalable solution suitable for deployment in various settings, including conversational agents and entertainment platforms.

5 Conclusion

This study has successfully demonstrated the advanced capabilities of the LLaMA-3-8B model, fine-tuned with LoRA in generating humor with high accuracy and relevance. By utilizing a robust dataset sourced from varied online platforms and non-humorous contents, alongside the employment of DistilBERT for efficient and reliable evaluation, our approach has set a new standard in computational humor generation. The results are exceptionally promising, with the LLaMA-3-8B model achieving an accuracy of 95.6% and an F1-score of 97.75%, surpassing larger models like GPT-4o and Gemini. These findings not only highlight the effectiveness of our data-driven and fine-tuning strategies but also underscore the potential of tailored computational models in understanding and generating humor.

The implementation of a comprehensive dataset and the strategic fine-tuning of the LLaMA-3-8B model have profound implications for the future of AI-driven humor generation. The study's focus on creating a discriminative model that effectively differentiates between humorous and non-humorous content paves the way for more sophisticated conversational agents and entertainment platforms. These agents can potentially deliver contextually appropriate humor that is both engaging and enjoyable, thereby enhancing user interaction and satisfaction.

Looking forward, the potential applications of this research are vast and socially impactful. The ability of AI to generate humor reliably and contextually can be particularly beneficial in therapeutic settings, such as providing companionship and emotional support to the elderly or individuals experiencing depression. By integrating the proposed model into conversational agents, these systems can be tailored to deliver personalized humor, which can play a significant role in improving mental health and well-being.

Further research could also explore expanding the dataset to cover a broader range of humor styles and cultural nuances, thereby enhancing the model's applicability and effectiveness across different demographics and geographical locations. Additionally, future work could focus on refining evaluation frameworks to incorporate more diverse metrics that capture subtleties in humor appreciation, addressing potential biases, and enhancing the model’s sensitivity to a wider array of humor types.

This research not only advances the field of computational humor in AI but also opens up new avenues for the practical application of humor-generating models in enhancing human-machine interaction and supporting mental health initiatives. The integration of refined AI models in user-facing applications promises not just to entertain but also to provide companionship and support, marking a significant step forward in the humane application of artificial intelligence.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

1. J. Porteous, Humor and social life. Philos. East West 39, 279–288 (1989).

2. S. Aaronson, Essentials of complexity-theoretic stand-up comedy. https://scottaaronson.blog/?p=414 (2009).
3. A. Cann, K. Stilwell, K. Taku, Humor styles, positive personality and health. Eur. J. Psychol. 6, 213–235 (2010).
4. W. Ruch, Psychology of humor, in V. Raskin (Ed.), The Primer of Humor Research (De Gruyter Mouton, Berlin, New York, 2008), pp. 17–100.
5. C. Larkin-Galiñanes, An overview of humor theory, in The Routledge Handbook of Language and Humor (Routledge, 2017), pp. 4–16.
6. S. Attardo, The semantic foundations of cognitive theories of humor. Humor 10, 395–420 (1997).
7. J. Wilkins, A.J. Eisenbraun, Humor theories and the physiological benefits of laughter. Holist. Nurs. Pract. 23, 349–354 (2009).
8. C. Warren, A. Barsky, A.P. McGraw, What makes things funny? An integrative review of the antecedents of laughter and amusement. Pers. Soc. Psychol. Rev. 25, 41–65 (2021).
9. A. Mendiburo-Seguel, D. Páez, F. Martínez-Sánchez, Humor styles and personality: A meta-analysis of the relation between humor styles and the Big Five personality traits. Scand. J. Psychol. 56, 335–340 (2015).
10. S. Jentzsch, K. Kersting, ChatGPT is fun, but it is not funny! Humor is still challenging Large Language Models. arXiv:2306.04563 (2023).
11. V. Sanh, DistilBERT, A Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. arXiv:1910.01108 (2019).
12. Conference on Fairness, Accountability, and Transparency, 1622–1636 (2022).
13. I. Annamoradnejad, G. Zoghi, Colbert: Using BERT sentence embedding in parallel neural networks for computational humor. arXiv:2004.12765 (2020).
14. J. Zhang, L. Jain, Y. Guo, J. Chen, K.L. Zhou, S. Suresh, et al., Humor in AI: Massive Scale Crowd-Sourced Preferences and Benchmarks for Cartoon Captioning. arXiv:2406.10522 (2024).
15. Y. Su, Computer-generated Humour Based on GPT-2, in Proceedings of the 2022 IEEE 2nd International Conference on Data Science and Computer Application (IEEE, 2022), pp. 890–893.
16. Z. Li, K. Ren, X. Jiang, B. Li, H. Zhang, D. Li, Domain generalization using pretrained models without fine-tuning. arXiv:2203.04600 (2022).
17. H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.A. Lachaux, T. Lacroix, et al., Llama: Open and efficient foundation language models. arXiv:2302.13971 (2023).
18. I. Kant, Kritik der Urteilskraft, Vol. 39, Meiner (1913).
19. D. Cortiz, Exploring transformers in emotion recognition: A comparison of BERT, DistilBERT, RoBERTa, XLNet and Electra. arXiv:2104.02041 (2021).
20. H. Qin, M. He, H. Jia, HIMA-Net: humor prediction by self-attention based on key information related to humor, in Proceedings of the 2021 International Conference on Neural Networks, Information and Communication Engineering (SPIE, 2021), pp. 157–163.
21. T. Chaudhary, M. Goel, R. Mamidi, Towards conversational humor analysis and design. arXiv:2103.00536 (2021).

22. Z. Han, J. Wu, C. Huang, Q. Huang, M. Zhao, A review on sentiment discovery and analysis of educational big-data. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 10, e1328 (2020).
23. H. He, E.A. Garcia, Learning from imbalanced data. IEEE Trans. Knowl. Data Eng. 21, 1263–1284 (2009).
24. H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.A. Lachaux, T. Lacroix, et al., Llama: Open and efficient foundation language models. arXiv:2302.13971 (2023).
25. J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, et al., Chain-of-thought prompting elicits reasoning in large language models, in Proceedings of the Advances in Neural Information Processing Systems 35 (2022), pp. 24824–24837.
26. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 15, 1929–1958 (2014).
27. A. Krogh, J. Hertz, A simple weight decay can improve generalization, in Proceedings of the Advances in Neural Information Processing Systems 4 (1991).
28. R. Pascanu, On the difficulty of training recurrent neural networks. arXiv:1211.5063 (2013).
29. M. Abadi, A. Chu, I. Goodfellow, H.B. McMahan, I. Mironov, K. Talwar, et al., Deep learning with differential privacy, in Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (ACM, 2016), pp. 308–318.
30. M. Abadi, A. Chu, I. Goodfellow, H.B. McMahan, I. Mironov, K. Talwar, et al., Deep learning with differential privacy, in Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (ACM, 2016), pp. 308–318.
31. T.B. Brown, Language models are few-shot learners. arXiv:2005.14165 (2020).
32. J. Kaplan, S. McCandlish, T. Henighan, T.B. Brown, B. Chess, R. Child, et al., Scaling laws for neural language models. arXiv:2001.08361 (2020).