Laughing Out Loud – Exploring AI-Generated and Human-Generated Humor

Hayastan Avetisyan*, Parisa Safikhani*, and David Broneske

Department of Research Infrastructure and Methods, DZHW, Hannover, Germany

Abstract. In this study, we conduct a thorough comparative analysis between artificial intelligence (AI)-generated humor and human humor. The objective is to acquire a more profound understanding of AI's present capabilities in generating humorous text. We investigate the structural, sentiment, and linguistic patterns in jokes created by AI and humans, evaluating 'funniness' and 'originality' via a comprehensive annotation process. Our findings indicate that AI can produce humorous and occasionally novel content. Additionally, we employed the RoBERTa model for humor detection on a dataset consisting of 500 entries, including both human and AI-generated humor. This model demonstrated its proficiency in accurately categorizing a large dataset encompassing up to 200,000 entries with remarkable accuracy of up to 98%. Nonetheless, it lacks the emotional depth and originality commonly seen in human humor. The study underscores the challenge involved in developing AI models that can generate humor equivalent to human communication. Future research should focus on enhancing AI's ability to create humor and further examine AI's potential to adopt human humor strategies. Despite some limitations, this study contributes significantly to improving the humorous capabilities of AI models and the expandability of AI-generated humor.

Keywords: Artificial Intelligence (AI)-generated humor, Human humor, Linguistic patterns, Funniness evaluation, Originality evaluation, RoBERTa model.

1 Introduction

Humor, an inherent aspect of human communication, offers a powerful tool for establishing connections, lightening the atmosphere, and conveying intricate messages. The ubiquity of humor in our everyday lives contrasts sharply with the complexities it presents within the field of Natural Language Processing (NLP), especially when it comes to the generation of AI-based humor.

With the advent of sophisticated AI technologies, we have seen machines produce text that approximates human humor. However, the true measure of quality and originality in this machine-generated humor requires further comprehensive examination. Most existing research restricts its focus to certain types of joke structures, such as puns or knock-knock jokes, thus offering a limited perspective on the broader capabilities and limitations of AI in humor generation. Additionally, there is a noticeable gap in research providing a holistic view of the differences and similarities between AI and human humor.

Addressing these shortcomings, our study embarks on an exhaustive comparison of AI-generated humor with that generated by humans. Drawing from two distinct datasets – one composed by a state-of-the-art language model (AI-generated), and the other by humans – we scrutinize the structure, sentiment, and linguistic patterns present in both types of humor.

An integral part of our research is dedicated to pinpointing the differences and similarities between AI and human humor. This comparative analysis illuminates areas where AI falls short, providing insight into the current limitations and possible pathways for enhancement. Conversely, the identified similarities serve as evidence of AI's success in

David C. Wyld et al. (Eds): ICCSEA, NLAI, SCAI, CSIA, IBCOM, SEMIT, NECO, SPPR, MLDS -2023
pp. 59-76, 2023. CS & IT - CSCP 2023 DOI: 10.5121/csit.2023.132406

emulating certain facets of human humor, a crucial step in the ongoing refinement of AI's humor generation abilities.

Further, we present a series of recommendations for future research based on our findings. As AI continues to advance and evolve, it becomes increasingly important to reassess and fine-tune its capabilities in humor generation. Our study suggests specific focus areas for future research, particularly the exploration of AI's potential to generate humor that is more original and emotionally resonant, striving towards a more human-like approach.

The contributions of this paper can be categorized into several key domains:

- **Comparing AI and human-generated humor**: Our comprehensive comparison of structure, sentiment, and linguistic patterns in AI and human-generated humor deepens the understanding of AI's current abilities in humor generation.
- **Evaluation of 'funniness' and 'originality'**: We thoroughly evaluate 'funniness' and 'originality', offering new insights into AI's proficiency in humor generation.
- **Human-annotated dataset**: A unique aspect of our work is the creation of an annotated dataset, in which both AI and human-generated humor have been evaluated by human annotators for 'funniness' and 'originality'. This dataset, publicly available for further research, provides a valuable resource for studying and understanding the attributes of humor as perceived by humans.
- **Identification of differences and similarities**: By identifying key differences and similarities between AI and human humor, we provide a greater understanding of AI's current limitations and potential areas for improvement.
- **Recommendations for future research**: Our study provides valuable guidance for future research, focusing on the enhancement of AI's capability to generate humor that is both original and emotionally resonant.

In summary, this paper significantly contributes to the understanding and improvement of AI's ability to generate humor, while highlighting the intricate nature of humor generation. The introduction of our annotated dataset adds a tangible dimension to the study, allowing for further, detailed exploration into human perception of humor in comparison to AI-generated humor.

2 Related works

2.1 Automated humor detection

In a recent study [1], researchers performed a comparative evaluation of several machine learning (e.g., logistic regression, decision tree, random forests, passive-aggressive classifier) and deep learning (e.g., CNN, LSTM) methods to classify tweets as either humorous or non-humorous. They used a Kaggle dataset comprising both types of tweets. The findings from this study indicated that deep learning techniques delivered superior accuracy in humor prediction when compared to traditional machine learning approaches. [2] introduced a deep learning-based humor detection technique that combined CNN and LSTM layers. This approach addressed CNN's contextual limitations using LSTMs and incorporated dropout layers to reduce overfitting. When tested on the Yelp user review dataset, the model outperformed other methods, including SGD, SVM, and XGBoost, in precision, recall, and F1-measure. The research hints at the technique's broader applications, even in areas like detecting psychopathic behavior on social platforms.

In another study[3], machine learning models were developed to identify and score humor and offense in text. Using a dataset of 8,000 sentences, the study compared BERT-based models (BERTBASE, DistillBERT, and RoBERTa) for high-performance humor detection. DistillBERT performed best for humor detection and rating, RoBERTa excelled in controversial detection, and BERTBASE outperformed in offensiveness ranking.

A recent study by [4] reviewed the field of computational humor recognition. They examined 106 papers from various databases and analyzed datasets, features, and algorithms used in the field. The study found numerous annotated humor datasets and identified 21 frequently studied humor features. The researchers observed that deep learning and supervised learning techniques, particularly Support Vector Machine and Long Short-Term Memory Networks, were commonly used for humor classification. BERT was the most popular pre-trained language model in this context. Future research directions and challenges in humor detection were also discussed.

2.2 AI- or human-generated text

In a recent study [5], researchers examined ChatGPT-generated online reviews. They compared human and ChatGPT content using a Transformer-based model with SHAP for explainability. Results showed a 79% accuracy rate, and they observed that ChatGPT tends to produce polite, vague, and impersonal text with unique vocabulary and minimal emotional expression.

Another paper [6] discusses the challenges of detecting AI-generated text and the potential risks of unregulated use of large language models (LLMs). The authors argue that current detectors are unreliable due to paraphrasing attacks and detector limitations. They also demonstrate the vulnerability of watermarked LLMs to spoofing attacks. The paper emphasizes the need for secure methods to prevent LLM misuse and the risks associated with misidentification by AI text detectors.

[7] present DetectGPT, a zero-shot machine text detection technique using the negative curvature regions of a language model's log probability function. It outperforms others without needing a separate classifier or dataset, using log probabilities and random perturbation for sample detection. Limitations include its white-box assumption, reliance on a sound perturbation function, and high computational needs. Future research could focus on watermarking, model ensembles, prompt-detection relationship, and applying this method to other domains.

The research by [8] explores distinguishing AI-generated text from human-written ones. It argues that detection is feasible unless both text distributions match. It presents a sample complexity bound for detection, stating more human-like AI text requires more samples. The study confirms better detectors are achievable, with more samples and robust watermarking techniques improving detection even amidst paraphrasing attacks. It highlights the need for further research for effective and fair AI text detectors.

[9] focuses on the differences between AI-generated and human-written scientific texts, exploring the potential limitations and challenges of using AI-based writing assistants in scientific writing. The researchers collected and analyzed scientific text from OpenAI API using optimized prompts for structured scientific abstracts. They conducted human evaluations to assess the ability to distinguish between AI-generated and human-written texts and devised a feature description framework to analyze differences in syntax, semantics, and pragmatics. Using logistic regression models and fine-tuned RoBERTa large OpenAI detectors, they found that AI-generated scientific texts distinguish from human-written texts, often lacking valuable insights and showing low external inconsistency with actual scientific knowledge. While AI-generated text may eventually become more syntactically similar to human-written text, future research should focus on improving the semantics and pragmatics of AI-generated texts to enhance human-AI collaboration in the research process. [10] examines the threat models posed by contemporary natural language generation (NLG)

systems and reviews the most complete set of machine-generated text detection methods available. With powerful open-source generative models becoming increasingly accessible, detecting machine-generated text is crucial in mitigating the potential for abuse. However, detecting machine-generated text presents several technical challenges and open problems. The paper provides a comprehensive analysis of the threat models posed by contemporary NLG systems and emphasizes the urgent need for improved defenses against the abuse of NLG models. The paper also highlights the importance of coordinated efforts across technical and social domains to achieve practical solutions.

Based on the review of related work, it becomes evident that while there have been numerous studies focusing on humor detection in text and distinguishing between AI-generated and human-generated text, no specific research has yet been undertaken to concentrate solely on the detection of humor generated by AI and contrasting it with human-generated humor. This existing gap in the literature motivates our current study.

3 Methodology

This chapter provides a comprehensive overview of the methodology adopted in our research to analyze and compare AI and human-generated humor effectively. It elaborates on three key aspects: the data set utilized, the manual annotation of the dataset, and the experiments conducted.

Our methodology's ultimate goal is to support the findings of this study through a transparent, replicable, and robust approach that not only substantiates our research outcomes but also serves as a guideline for future research in this intriguing domain.

3.1 Dataset

AI-generated dataset. We utilized ChatGPT (GPT-4) to create an AI-generated dataset for humor detection, which comprised both humorous and non-humorous entries. We directed the model with the prompt: "Provide a balanced dataset of 500 entries for humor detection." Recognizing ChatGPT's limitation in generating 500 entries in a single go, we intermittently prompted the model every 20 entries using the prompt "provide more entries" to continue the generation process. Upon reviewing the first batch of 500 generated entries, we noted that while the dataset was balanced, there were significant duplicates in the humorous section. To obtain a well-balanced dataset free from duplicates, we had to prompt the model a total of 970 times.

Human-generated dataset. The Colbert dataset[1] is an assortment of textual data that was employed to examine the application of BERT (Bidirectional Encoder Representations from Transformers) sentence embeddings in the detection of humor. This dataset is well-balanced, as indicated by [12]. The dataset has 200.000 entries, but to have a fair comparison with the ChatGPT-generated dataset, we have randomly reduced it to 500 entries while keeping the proportions of humorous and non-humorous labels stable.

Test set. The human-generated test set comprises 100 meticulously selected entries from Colbert's content. These entries were randomly selected to capture a representative sample of Colbert's unique speech patterns and nuances. Extra care was taken to ensure that these entries did not overlap with any from the training set.

The AI-generated test set was produced by Bing ChatBot. The same prompting methodology was employed when interacting with Bing ChatBot. We chose Bing ChatBot

[1] `https://www.kaggle.com/datasets/deepcontractor/200k-short-texts-for-humor-detection`

for testing because it offers a valuable external benchmark for our models, which are trained on the generated text by the ChatGPT model. Not only does it allow us to compare performance against a distinct AI, but it also helps us understand how our model interacts with diverse dialogue styles.

3.2 Dataset annotation

The task of annotating our datasets was a pivotal step in our research, given the subjective nature of humor and its dependence on various nuanced factors. Our annotation scheme was designed to help the annotators rate the humor of jokes or humorous texts based on 1) funniness and 2) originality. This section will provide a detailed overview of this annotation process.

Two human evaluators annotated the AI-generated and human-generated humor datasets. They were given clear instructions about the task and were blind to the source of the texts (AI or human) to prevent any bias.

The annotators rated the **’funniness’** of the text on a Likert scale ranging from 1 to 5, where 1 represented ’not funny at all’ and 5 meant ’hilarious’. The intermediate values 2, 3, and 4 indicated increasing levels of funniness.

The **’originality’** of the humor was also rated on a Likert scale from 1 to 5, where 1 indicated ’not original at all,’ and 5 showed ’highly original’. The intermediate values represented increasing levels of originality.

Annotators were provided with clear guidelines to maintain objectivity, consider cultural context, ensure a complete understanding of the text, and maintain consistency in their ratings.

The annotated data then underwent a cleaning process, where we resolved any disputes in ratings through discussions or by referring to the opinion of a third evaluator. This process resulted in a robustly annotated dataset that laid the groundwork for our comparative analysis of human and AI-generated humor.

4 Experiments

To delve deeper into the intricacies of humor produced by both humans and artificial intelligence (AI), we conducted a series of experiments using the RoBERTa model, which is the most suitable model for humor detection [21]. Utilizing our provided dataset, we aimed to determine how RoBERTa can distinguish between humor in human-generated and AI-generated texts, as well as discern humor generated by AI from that created by humans[2].

4.1 Experiment 1

In our initial experiment, we fine-tuned the RoBERTa model for the humor detection task using two specific datasets: Human-Generated Humor Detection and AI-Generated Humor Detection. By using a learning rate of 5e-5, a maximum sequence length of 128, and implementing the Adam optimizer, we prepared the model for higher performance. The efficiency of each model was tested against corresponding datasets, i.e., the AI-tuned model was assessed with the AI_test set and the human-tuned model with the Human_test set.

[2] `https://github.com/DZHW-AI4SS/Laughing-Out-Loud-Exploring-AI-Generated-and-Human-Generated-Humor.git`

4.2 Experiment 2

Our second experiment aimed at tuning the RoBERTa model to discern whether any given text, humorous or otherwise, was generated by a human or an AI system. For this, we utilized a hybrid dataset containing both AI-produced and human-produced texts, labeled '1' and '0' respectively. We gauged the model's accuracy with a specially compiled test set featuring both AI and human-generated content in a humor detection scenario.

4.3 Experiment 3

To further explore the variances between human and AI-generated humor, we modified our dataset by removing the non-humorous components. After this adjustment, we repeated the second experiment, focusing solely on the humorous content generated by both human and AI sources in order to detect whether there is a difference between AI-generated and human-generated humor.

5 Results and discussion

5.1 Overview of experimental results

According to the results of the first experiment presented in Table 1, the RoBERTa model fine-tuned on both datasets reached perfect F1 scores, precision, and recall of 100%. These results indicate an impeccable performance of the models on both the Human and AI test datasets, implying that the models were able to correctly detect all instances of humor without any false positives or false negatives. It is interesting to note that the model fine-tuned on the AI-generated humor detection dataset reached this level of accuracy more rapidly, achieving perfect scores in just 2 epochs, compared to the 4 epochs required for the Human dataset. This suggests that the complexity of human-generated humor is higher than in AI-generated humor.

Source	F1 Score	Precision	Recall	Epoch
Human	100%	100%	100%	4
AI	100%	100%	100%	2

Table 1. Comparison of the performance of RoBERTa fine-tuned on Human- and AI-generated Humor detection datasets

The results of second experiment are shown in Table 2. In the initial phase, we tested the model on a dataset that encompassed both humorous and non-humorous content generated by humans and an AI (Bing chatbot). The results presented in Table 2 demonstrated the model's extraordinary performance with perfect scores of 100% in F1, precision, and recall metrics, reached within 4 epochs. This indicates the model was successful in precisely classifying whether the text was produced by a human or an AI, regardless of whether the

Source	F1 Score	Precision	Recall	Epoch
AI/Human	100%	100%	100%	4

Table 2. Performance of the fine-tuned RoBERTa model, tuned to AI- or human-generated labels, on a test set generated by a human and a Bing chatbot with humorous and non-humorous context.

text was humorous or not. To examine the model's generalizability, we applied it to the extensive Colbert dataset. Impressively, the model accurately classified the entire dataset as human-generated text, achieving a 98% accuracy rate. These results underscore the potential of using a small, diverse dataset of AI and human-generated humor to construct a model that can efficiently and accurately categorize large-scale data, even when reaching up to 200,000 entries.

The outcomes of the third experiment are presented in Table 3. This phase concentrated solely on humorous text produced by both humans and AI. The results from this phase, as shown in Table 3, indicated an F1 score of 97.99%, precision of 98.07%, and recall of 98%, all reached within 2 epochs. These results suggest that while the model was slightly less accurate in identifying the source of humorous content compared to the mixed content, it still performed impressively.

Source	F1 Score	Precision	Recall	Epoch
AI/Human	97.99%	98.07%	98%	2

Table 3. Performance of the fine-tuned RoBERTa model, tuned to AI- or human-generated labels, on a test set generated by a human and a Bing chatbot with just humorous context.

These outcomes reaffirm that the RoBERTa model, when appropriately fine-tuned, can effectively discern between human- and AI-generated text, even within the complex domain of humor. However, it is notable that the performance slightly decreases when the model is solely exposed to humorous content, indicating the potentially increased complexity or variability in the way humans and AI generate humor.

5.2 Dataset analysis findings

Comparative analysis of AI-generated and human-generated humor This subchapter provides a comparative examination of the linguistic patterns [17–20] discerned in the AI-generated and human-generated humor datasets. The features analyzed in this study, including bigram usage, sentiment distribution, Part-of-Speech (POS) distribution, and average text length, were selected to provide a comprehensive understanding of the linguistic aspects of humor generation. These features allow us to explore the specific linguistic patterns, emotional tones, and linguistic elements employed by AI and humans in generating humor. By examining these features, we gain insights into the mechanisms and strategies behind humor generation, contributing to a deeper understanding of the similarities and differences between AI and human-generated humor.

Bigram usage: The top 10 bigrams in both AI-generated and human-generated humor datasets demonstrated considerable overlap. Common bigrams, such as 'do, you', 'what, do', 'you, call', 'call, a', 'did, they', 'in, the', 'what, the', 'what, is', and 'is, a', and 'knock, knock' (see Figures 1, 2), were prevalent in both datasets, indicating similarities in language structure used in the context of humor generation. However, a more detailed analysis should be conducted in future work, which might reveal slight variations in frequency and usage context.

Sentiment distribution: The sentiment distributions in both AI-generated and human-generated datasets indicate significant similarities, suggesting that the AI model has generally succeeded in capturing the emotional nuances of human humor (refer to Figures 3, 4). However, upon closer inspection, distinct differences emerge. The AI tends to produce humorous texts that are more neutral compared to those created by humans. Human-generated texts exhibit broader sentiments, some veering towards positive or negative tones.

Despite this, the overall sentiment in both groups remains low, with most texts falling into the neutral category.

Part-of-Speech (POS) distribution: Both datasets presented a similar distribution of POS tags, with nouns, punctuations, pronouns, determiners, verbs, and auxiliaries (see Figures 5, 6) being the most commonly used. This similarity might suggest that the AI has effectively mirrored human syntactic structures when generating humor. Nonetheless, a deeper inspection might uncover subtle differences in the context or complexity of usage across POS categories.

Average number of words: The average number of words in the AI-generated humor texts was found to be slightly lower than in the human-generated ones (12 compared to 14) (see Figures 7, 8). This difference indicates that while the AI generates humor within a relatively comparable length, it tends to create slightly more concise content. The reason for this verbosity disparity could be an interesting point of further investigation.

This comparative analysis serves as a stepping stone towards a deeper understanding of the differences and similarities between human and AI humor generation. It provides crucial insights that will guide the subsequent steps of our investigation and the interpretation of our experimental results.

5.3 Annotation analysis findings

The annotation process played a crucial role in evaluating the quality of humor in both AI-generated and human-generated texts. Two critical parameters assessed during annotation were ’funniness’ and ’originality’. The following subchapters delve into the findings for each of these categories.

Funniness:
Upon closer inspection of the ’funniness’ ratings between both AI and human-generated humor (see Figure 9) the following trends emerge:

”Not at all funny” (Rating 1): Human-generated humor was rated as ”not at all funny” more frequently than AI-generated humor. This suggests that there are instances where human humor fails to resonate or generate amusement, at least more often than AI.

”Somewhat funny” (Rating 2): Both AI and human-generated humor were often perceived as somewhat funny, with human humor taking a slight lead. This indicates a shared capacity to generate humor that contains minor amusing elements but is unlikely to stimulate laughter.

”Moderately funny” (Rating 3): The frequency of this rating is similar for both AI and human humor, suggesting that both can produce content that is amusing to some extent and could potentially incite laughter.

”Very funny” (Rating 4): Surprisingly, AI-generated humor surpassed human humor in achieving the ”very funny” rating. This shows that the AI could generate humor that is quite amusing and likely to elicit laughter more often than humans.

”Hilarious” (Rating 5): Neither AI nor human humor managed to achieve the highest rating of being ”hilarious”. This suggests a common challenge in generating extremely amusing humor and sure to cause laughter.

These findings illustrate that while AI can generate amusing content, with a surprising lead in developing ”very funny” content, it still has room for improvement. The fact that both AI and humans struggled to achieve the ”hilarious” rating underlines the complexity of crafting universally appealing and highly effective humor. It emphasizes the importance of further refinement in AI’s humor generation techniques and the potential benefits of learning from human humor generation to enhance performance.

Originality:
In the analysis of the ”originality” ratings for both AI and human-generated humor (see Figure 10), we identify clear trends that provide insight into the novelty of humor produced by both entities.

The ratings suggest that human-generated humor is perceived as more original than AI-generated humor in most cases. This is particularly evident in the frequency of the ratings **”moderately original” (Rating 3)** and **”very original” (Rating 4)** for human humor, which far surpass those for AI. It shows that humans more frequently than AI combine original ideas with common joke structures or rely on unique, new ideas and unexpected twists to elicit laughter.

In contrast, the humor generated by the AI was most frequently rated as **”somewhat original” (Rating 2)**, indicating that the AI can incorporate some new elements but relies heavily on known jokes or structures. Additionally, the rating of **”not at all original” (Rating 1)** suggests that a portion of the AI humor is perceived as hackneyed, heavily recycled, or unoriginal, to a greater extent than human-generated humor.

Like the ”funniness” ratings, there were no ratings of **”extremely original” (Rating 5)** in both datasets, indicating humor that is unique, creative, and unlike anything seen before. This again underscores the shared challenge of creating humor that blazes new trails altogether.

These patterns highlight the current limitations of AI in generating humor that is seen as original and novel. They underline the need for AI to learn more from human humor, especially incorporating unique ideas and unexpected twists. Despite AI’s ability to generate humor with some degree of novelty, there remains a significant gap between the AI and human capacity to produce humor seen as highly original.

In conclusion, while AI has demonstrated an ability to generate humor with some original elements, it is still far from matching human capacity in originality. The findings suggest there is substantial room for AI to improve its performance in this area, particularly by learning from the human ability to create humor that blends original ideas with common joke structures or bases humor on unique, new ideas and unexpected twists.

5.4 AI vs. human humor: key differences

The comparative analysis of AI and human-generated humor provides crucial insights into NLP’s current state of humor generation. Based on the results, several key differences and similarities in the humor of both entities can be identified.

First, concerning the structural aspects of humor generation, both AI and humans show similar use of bigram combinations and part-of-speech (POS) tags. These similarities suggest some success of the AI model in mimicking human language structures in the context of humor generation. However, the AI’s tendency to produce more neutral sentiments than humans indicates potential gaps in the model’s understanding of emotional nuance, an essential aspect of humor. In addition, the AI tended to produce more concise content, indicating a difference in verbosity.

In terms of ’funniness’, both humans and AI have difficulty consistently generating highly amusing or hilarious content, illustrating the complexity of humor generation. Interestingly, AI humor received a ’very funny’ rating more often than human humor. This suggests that despite its limitations, the AI model can generate content that could elicit significant hilarity.

When analyzing the ’originality’ of the humor generated, human-generated humor is considered more original in most cases. This indicates that humans are more effective at combining original ideas with common joke structures or innovating with unexpected twists. In contrast, AI humor often relies heavily on familiar jokes or structures, resulting in less original humor.

These findings highlight the current limitations of AI and the areas where it can potentially learn from human humor generation. Despite its ability to generate humor with some novelty, a significant gap remains between AI's and humans' ability to generate highly original humor. The findings suggest that improvements in AI's understanding of emotional nuance, its ability to innovate beyond known joke structures, and its fusion of original and common elements could lead to more effective humor generation.

In summary, while AI can generate entertaining and reasonably original content, it still falls short of the originality and emotional depth often found in human-generated humor. The comparison reveals the potential areas of improvement for AI and the inherent complexity of humor generation, which presents a fascinating challenge for future research in NLP.

6 Conclusion and future work

This study presents a comprehensive comparative analysis between AI-generated and human-generated humor. It provides crucial insights into the structural, emotional, and linguistic patterns inherent in the humor generated by both. The ratings of 'funniness' and 'originality' that emerge from a thorough annotation process further illuminate the strengths and weaknesses of humor generation by AI and humans.

Our findings highlight that while AI shows promise in generating entertaining and sometimes original content, it does not achieve the emotional depth and originality often seen in human humor. The AI's tendency to create neutral moods, its reliance on familiar joke structures, and its difficulty in consistently generating highly entertaining content highlight potential areas for improvement.

In contrast, despite its perceived weaknesses in generating 'hilarious' content, human humor offers a broader range of moods and more original content. These findings serve as a reminder of the complex nuances involved in humor generation and the subtleties of human communication that AI has yet to grasp fully.

Looking forward, several promising avenues for further work emerge. First, investigating AI's understanding and use of different joke structures could provide insights into the effectiveness of these structures in humor generation. Second, more work is needed to improve AI's ability to innovate beyond known joke structures and incorporate original elements. Last, training AI models on more diverse and extensive humor datasets could help to better capture the variability and depth of humor in human communication.

These studies are expected to refine AI's humor generation capabilities and enrich our understanding of the complex nature of humor. As we continue to move forward in this fascinating intersection of AI and humor, we look forward to the day when machines can generate humor that matches, or perhaps surpasses, the hilarity and originality of human humor.

In summary, the quest to give AI the ability to generate high-quality, original humor remains a challenging but exciting journey. We hope that the findings of this study will stimulate further research and support ongoing efforts to improve the performance of AI in humor generation, and bring us one step closer to this elusive goal.

7 Limitations

This study, aimed at comparing AI-generated and human humor, offers significant insights but has several limitations that necessitate careful interpretation of the findings.

The first limitation pertains to the potential subjectivity in the 'funniness' and 'originality' ratings. Humor, a highly personal and culturally influenced trait, may have been perceived and rated differently by different evaluators. This subjective variation could have affected the perceived quality of humor across the AI and human-generated content.

Secondly, the generalizability of the findings may be restricted due to the specificity of the dataset used in the analysis. Given the complexity and diversity of humor, the study's findings might not apply to all forms of humor or other datasets. Therefore, we may have missed capturing the full spectrum of humor by restricting the analysis to a single dataset.

The third limitation lies in the study's focus. While the study thoroughly analyzes the structure and originality of humor, it does not consider the impact of specific types of content such as satire, irony, or dark humor. Due to their unique nature, these humor styles might necessitate a different analytical approach and could influence the interpretation of the results.

These limitations indicate the need for future research to expand the scope of analysis to include a wider variety of datasets and humor types and potentially employ multiple raters for a more comprehensive and balanced assessment. Despite these limitations, the study underscores an essential progression in the field and contributes towards enhancing the ability of AI models to generate humor.

References

1. Prajapati, Pariksha, et al. "Empirical Analysis of Humor Detection Using Deep Learning and Machine Learning on Kaggle Corpus." International Conference on Advancements in Interdisciplinary Research. Cham: Springer Nature Switzerland, 2022.
2. Kumar, Vijay, Ranjeet Walia, and Shivam Sharma. "DeepHumor: a novel deep learning framework for humor detection." Multimedia Tools and Applications 81.12 (2022): 16797-16812.
3. Mathias, Marcelo Custódio. Humor and offense speech classification and scoring using natural language processing. MS thesis. 2022.
4. Kalloniatis, Antony, and Panagiotis Adamidis. "Computational Humor Recognition: A Systematic Literature Review." (2023).
5. Mitrović, Sandra, Davide Andreoletti, and Omran Ayoub. "Chatgpt or human? detect and explain. explaining decisions of machine learning model for detecting short chatgpt-generated text." arXiv preprint arXiv:2301.13852 (2023).
6. Sankar Sadasivan, Vinu, et al. "Can AI-Generated Text be Reliably Detected?." arXiv e-prints (2023): arXiv-2303.
7. Mitchell, Eric, et al. "Detectgpt: Zero-shot machine-generated text detection using probability curvature." arXiv preprint arXiv:2301.11305 (2023).
8. Chakraborty, Souradip, et al. "On the possibilities of ai-generated text detection." arXiv preprint arXiv:2304.04736 (2023).
9. AI vs. Human–Differentiation Analysis of Scientific Content Generation.
10. Crothers, Evan, Nathalie Japkowicz, and Herna L. Viktor. "Machine-generated text: A comprehensive survey of threat models and detection methods." IEEE Access (2023).
11. Kumar, Vijay, Ranjeet Walia, and Shivam Sharma. "DeepHumor: a novel deep learning framework for humor detection." Multimedia Tools and Applications 81.12 (2022): 16797-16812.
12. Annamoradnejad, Issa, and Gohar Zoghi. "Colbert: Using bert sentence embedding for humor detection." arXiv preprint arXiv:2004.12765 1.3 (2020).
13. Elayan, Suzanne, et al. "'Are you having a laugh?': detecting humorous expressions on social media: an exploration of theory, current approaches and future work." International Journal of Information Technology and Management 21.1 (2022): 115-137.
14. Cao, Yihan, et al. "A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt." arXiv preprint arXiv:2303.04226 (2023).
15. Krishna, Kalpesh, et al. "Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense." arXiv preprint arXiv:2303.13408 (2023).
16. Li, Zhuohang, Jiashuo Liu, and Yuci Wang. "Performance Analysis on Deep Learning Models in Humor Detection Task." 2022 International Conference on Machine Learning and Knowledge Engineering (MLKE). IEEE, 2022.

17. Teller, Virginia. ”Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition.” (2000): 638-641.
18. Manning, Christopher, and Hinrich Schutze. Foundations of statistical natural language processing. MIT press, 1999.
19. Grishman, Ralph. Computational linguistics: an introduction. Cambridge University Press, 1986.
20. Eisenstein, Jacob. Introduction to natural language processing. MIT press, 2019.
21. Adoma, Acheampong Francisca, Nunoo-Mensah Henry, and Wenyu Chen. ”Comparative analyses of bert, roberta, distilbert, and xlnet for text-based emotion recognition.” 2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP). IEEE, 2020.

Hayastan Avetisyan completed her BA in Translation Studies in Yerevan, Armenia, and pursued an MA in Linguistics in Hannover, Germany, in 2020. In 2021, she joined the AI4S2 project at the Department of Research Infrastructure and Methods, focusing on NLP, ML, and AI interpretability. Her research explores leveraging linguistic knowledge to enhance the development and interpretation of language models. Currently pursuing her Ph.D., she investigates the utilization of AI in research methodologies.

Parisa Safikhani is a research scientist at the German Centre for Higher Education Research and Science Studies (DZHW). She holds an M.Sc. degree in electrical engineering and automation technology from LUH and obtained her Bachelor’s degree in electrical and telecommunication engineering from Arak University. Currently, she is pursuing her PhD in the field of artificial intelligence, with a specific focus on AutoNLP.

David Broneske is the head of the department Infrastructure and Methods at the German Centre for Higher Education Research and Science Studies (DZHW), Hannover. He received his PhD in Computer Science from the University of Magdeburg, where he also pursued his Master’s and Bachelor’s Studies in Computer Science. His research interests include main-memory database systems, interdisciplinary data management, and the application of artificial intelligence in various domains.

A Appendix

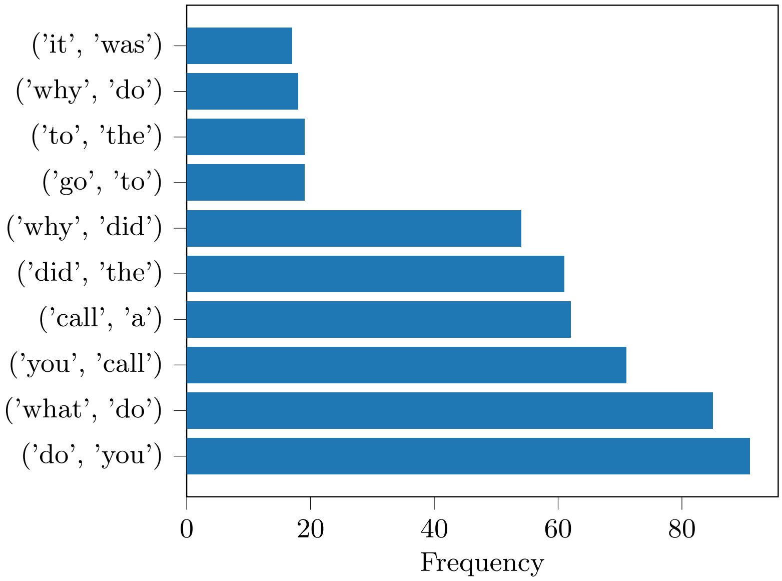

Fig. 1. Top 10 Bigrams of AI-generated humorous content.

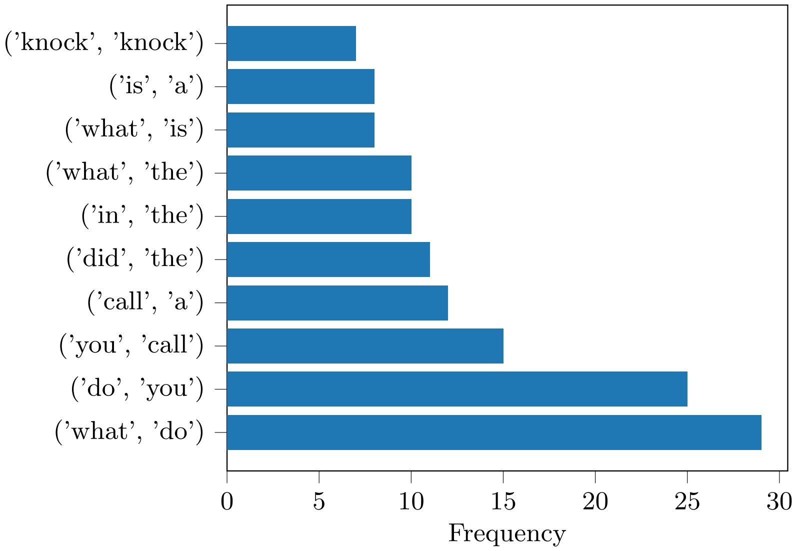

Fig. 2. Top 10 Bigrams of Human-generated humorous content.

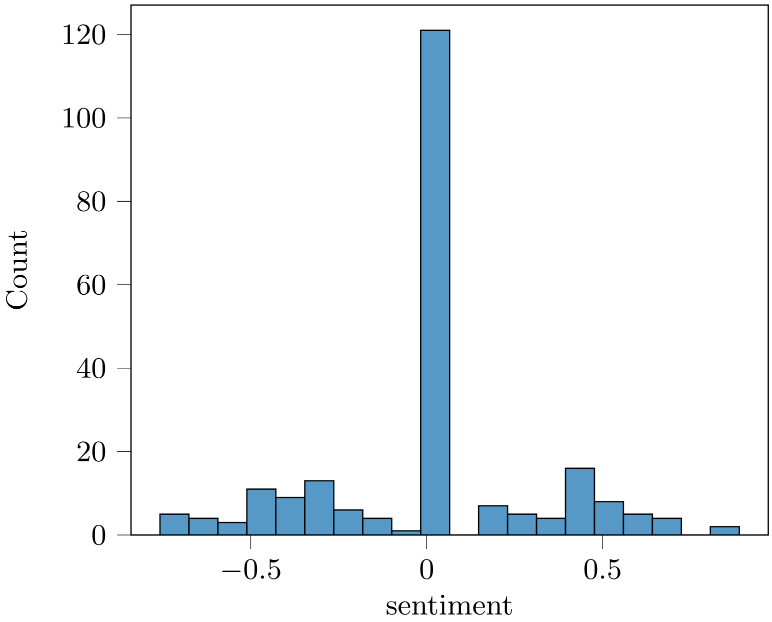

Fig. 3. Sentiment Distribution of AI-generated humorous content.

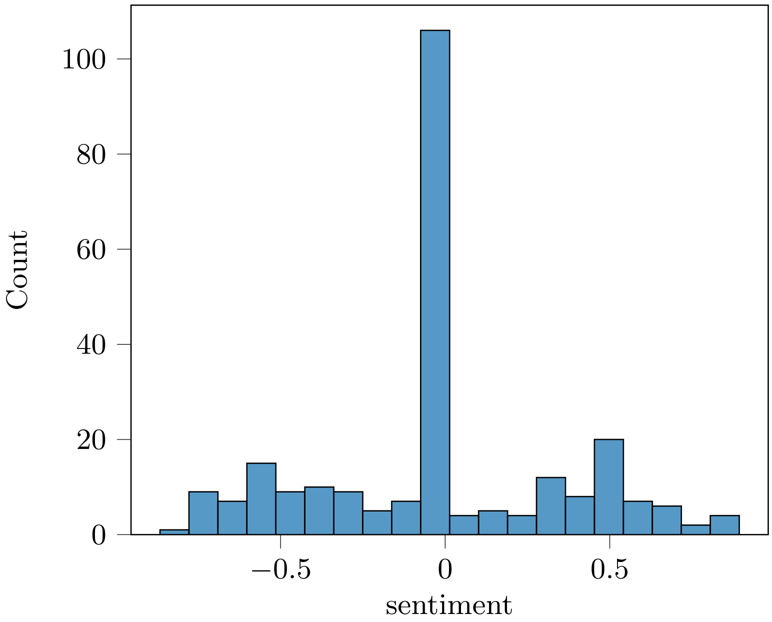

Fig. 4. Sentiment Distribution of Human-generated humorous content.

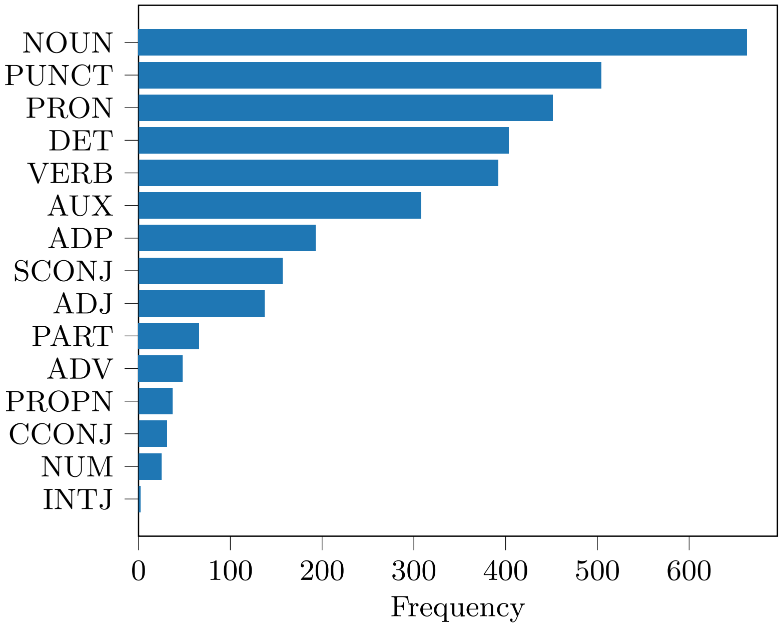

Fig. 5. POS Distribution of AI-generated humorous content.

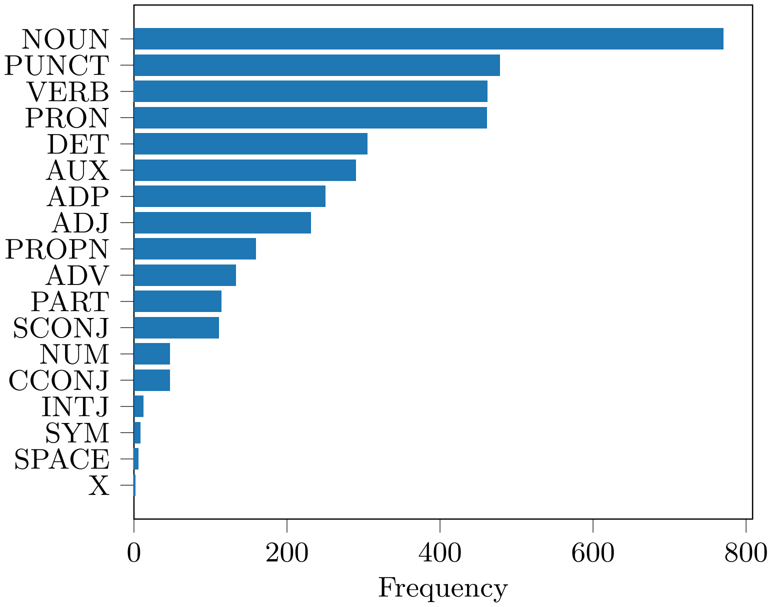

Fig. 6. POS Distribution of Human-generated humorous content.

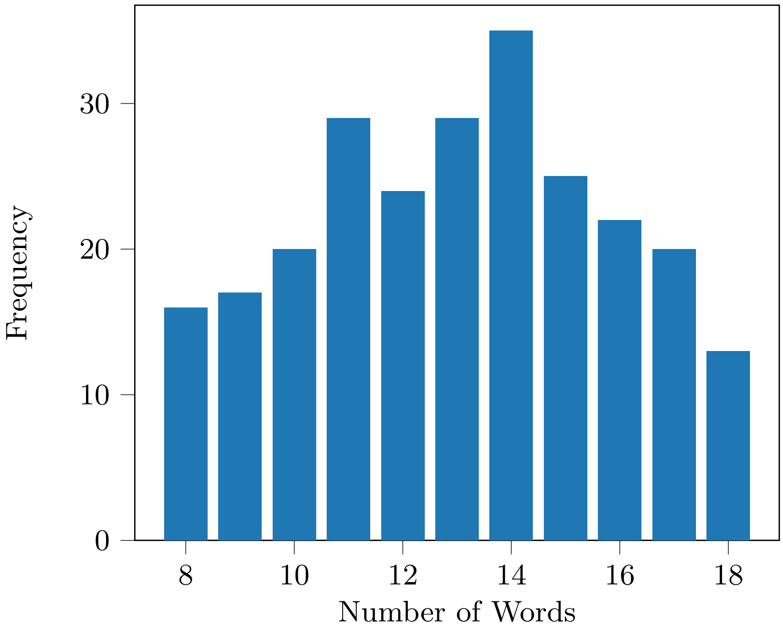

Fig. 7. Average number of words: Human-generated humorous content.

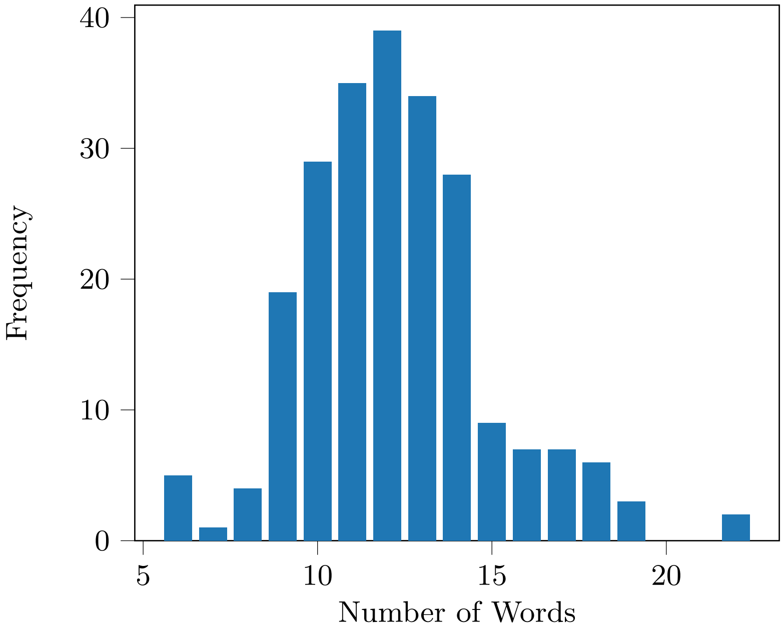

Fig. 8. Average number of words: AI-generated humorous content.

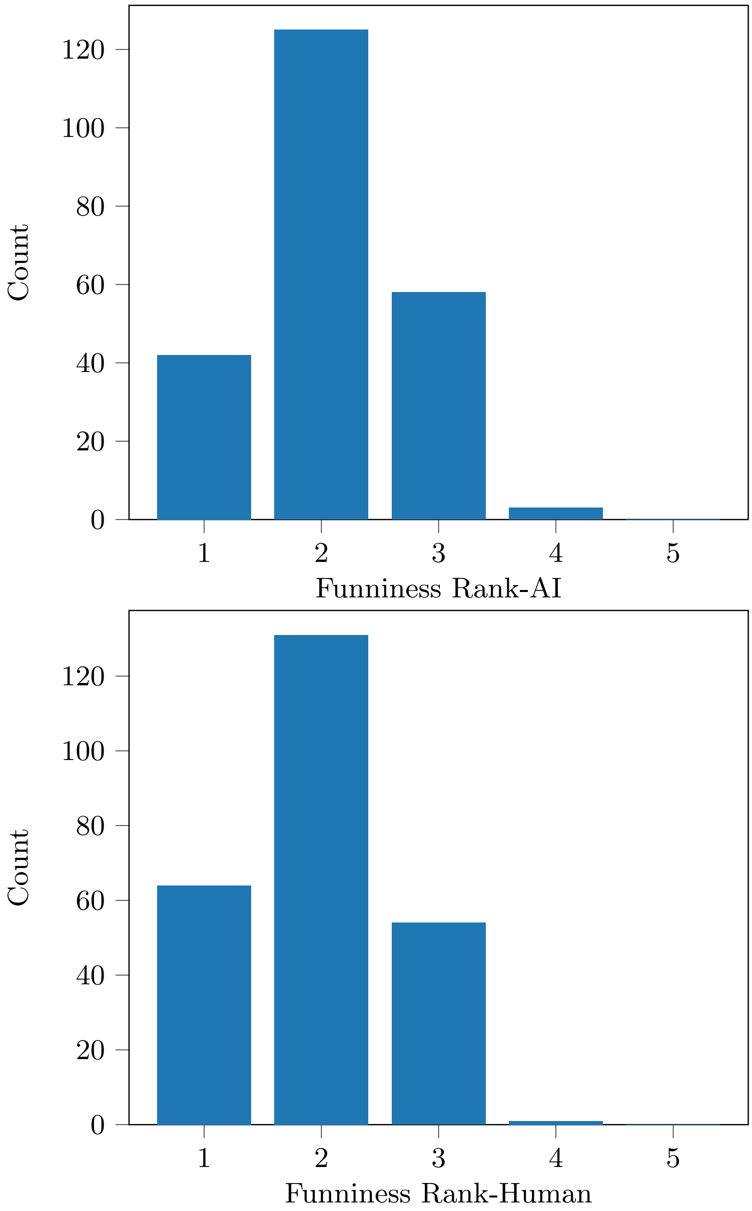

Fig. 9. Funniness of AI- and Human-generated humorous content.

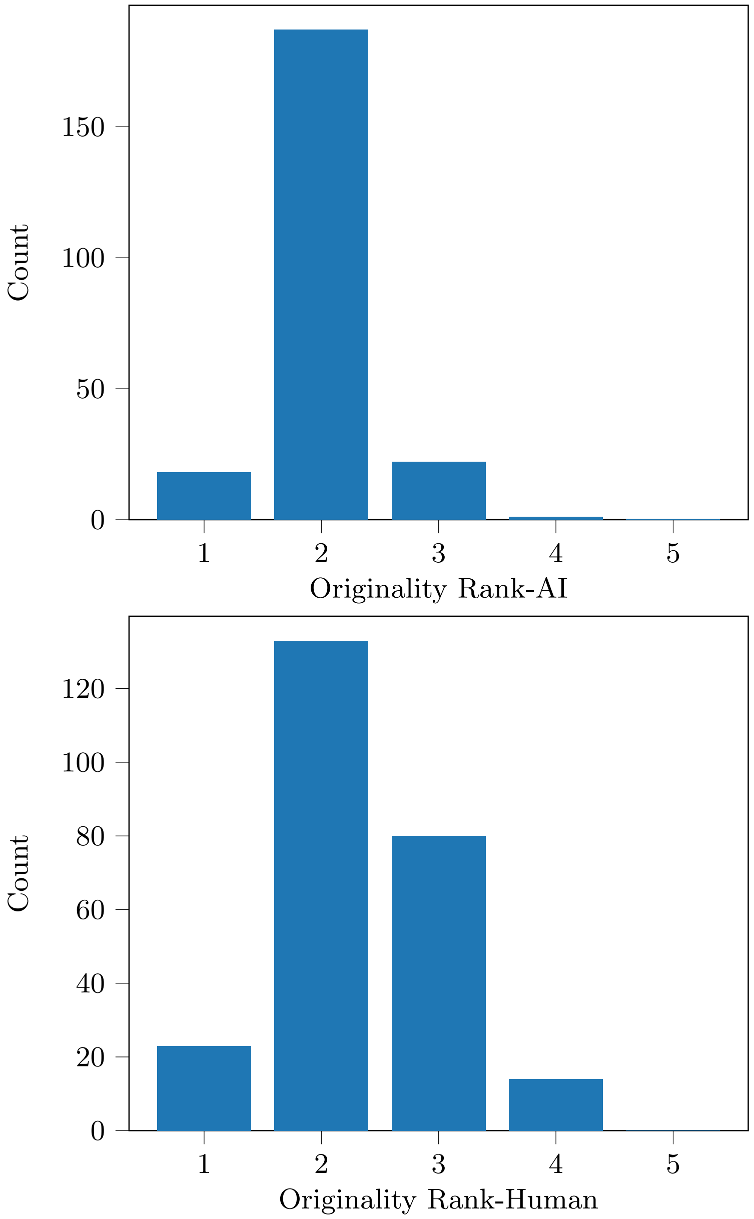

Fig. 10. Originality of AI- and Human-generated humorous content.