Title: "AITA Post Generator: A Computational Creativity System for Generating Authentic "Am I the A\*\*hole?" Posts"

Abstract:

In this project, I propose a computational creativity system for generating authentic "Am I the A\*\*hole?" (AITA) posts for the popular online forum. My goal with this system is to create engaging and believable content that sparks user discussions and fosters community interactions. I combine sentiment analysis, Proximal Policy Optimization (PPOs), and GPT to produce human-like content. The results of my system demonstrate its effectiveness in generating creative and contextually relevant AITA posts. I also discuss the ethical implications, challenges, and limitations of my project.

Introduction:

Online communities such as Reddit have become increasingly popular platforms for users to share experiences, engage in discussions, and seek advice from peers. One such community, the "Am I the A\*\*hole?" subreddit, allows users to post about their moral dilemmas and seek feedback from others. However, with a growing user base, maintaining engagement can be a challenge. Our project aims to address this challenge by developing an AITA post generator using computational creativity techniques, thereby fostering more engaging discussions and maintaining the vibrancy of the community.

Background:

Computational creativity has recently emerged as a promising field of research, with applications spanning various domains such as art, music, and natural language generation. Many studies have concentrated on generating contextually relevant content, drawing inspiration from state-of-the-art computational creativity systems such as OpenAI's GPT models and GAN-based text generation systems like StackGAN and AttnGAN Building on this research, the project specifically focuses on generating AITA posts, which necessitate a unique blend of sentiment, storytelling, and ethical consideration. By applying advances in computational creativity, I aimed to create a system capable of producing engaging and authentic content for the AITA community.

Methodology and Design:

My approach consists of three main components: dataset collection and preprocessing, model architecture, and training and evaluation methods.

Dataset Collection, Preprocessing, and Text Embedding: I obtained a dataset of over 100,000 popular and relevant AITA posts from the Kaggle repository (1), which had been scraped from the r/AmItheAsshole subreddit. I preprocessed the dataset to remove extraneous information, such as usernames and timestamps, and to standardize formatting. In the code, I first load the dataset and further preprocess it by dropping NaN values, balancing the classes, and concatenating the title and body of the posts. I then partition the data into training and test sets, and apply text preprocessing techniques, such as lemmatization and tokenization. Finally, I utilize GloVe embeddings to generate an embedding matrix for the preprocessed text.

Model Architecture: For my own generator, I made use of a slightly more advanced seq2seq model called multi-task learning for generating the AITA posts(2). The dataset consists of two classes, where each class represents whether the post's author is considered an a\*\*hole or not. I preprocessed the data, filtered it, and balanced the classes by sampling an equal number of instances from each class. The data is then tokenized, and the input sequences are padded to a fixed length.

The multi-task learning model is designed using an encoder-decoder architecture, with one encoder for processing the input sequences and two decoders for generating AITA titles and story bodies. The model is trained using the Adam optimizer and the sparse categorical cross-entropy loss function, with equal weights given to both title and body decoders. I chose to use the multi-task learning model instead of just a seq2seq model as I wanted the body to be related to the title. If I just trained a title generator and body generator separately, there is a high likelihood that what the title mentions and what the body produces would be entirely different.

After training, I used the decoder models to generate titles and bodies for AITA posts. To evaluate the generated text, I computed rewards based on the presence of keywords, sentiment, and grammar. The title reward is calculated based on the presence of title keywords and grammar quality, while the body reward is determined by the presence of situation, action, and justifiable keywords, as well as sentiment and grammar quality. The multi-task learning model is well-suited for text generation tasks due to its ability to handle variable-length input and output sequences.

To enhance the quality and context of the generated content, I incorporated a sentiment analysis module and Proximal Policy Optimization (PPO). PPO is a reinforcement learning algorithm that helps in fine-tuning the model to generate better AITA posts by optimizing the overall reward based on the evaluation criteria. The sentiment analysis module uses a pre-trained sentiment classifier to identify the emotional tone of the input text. This information is then used to guide the generation process, ensuring the generated posts convey appropriate emotions. In addition to this, I have also created an AITA classifier using a Bidirectional LSTM model with pre-trained GloVe embeddings to categorize posts into "Asshole" and "Not Asshole" classes. The Bidirectional LSTM model allows the system to capture both past and future context when making predictions, improving its classification performance.

For comparison purposes, I also created a version of our application that uses the GPT-Neo model as a benchmark to compare our system's performance against an advanced language model. I selected GPT-Neo over GPT-2 and GPT-3 for several reasons: it is more sophisticated than GPT-2, it is open-source, and it is not constrained by API requests like GPT-3. By leveraging the capabilities of GPT-Neo, i aimed to evaluate the quality of our generated AITA posts and see how well our system fares against state-of-the-art models.

Training and Evaluation Methods: I trained our seq2seq model using a combination of supervised learning, in which the model was trained on pairs of input and output text data, and unsupervised learning, where the model was fine-tuned using reinforcement learning techniques based on the feedback from the AITA classifier. The AITA classifier was trained on a balanced dataset with equal instances from both classes. i compared our results to existing computational creativity systems, such as GPT-2 and GAN-based text generation models. For the AITA classifier, i used metrics such as confusion matrix, accuracy score, and classification report to assess its performance.

Results:

AITA Classifier

I began the evaluation with the AITA post classifier, assessing its performance using accuracy, validation accuracy, loss, and validation loss. Initially, I had a dataset of 100,000 posts, but due to limited computing resources, I scaled it back to a smaller size. Using a confusion matrix, i achieved a classification rate of around 60%. Admittedly, i was not satisfied with these results at first, and attempted to improve the model by incorporating regularization and early stopping to prevent overfitting.

However, after researching ways to enhance our classifier's performance, I came across the original dataset authors' work, which reported a 62.0% accuracy using logistic regression with 1-gram frequencies as features. This result was achieved with a more extensive dataset and a stronger classifier, the BERT architecture, which only achieved a 61% accuracy on held-out data. Considering that my Bidirectional LSTM model is relatively simple compared to BERT and that i used a smaller dataset, our classifier's performance was deemed satisfactory given these constraints.

It is important to note that the AITA classification task is inherently subjective, which posed additional challenges during training. Contextual understanding is crucial for determining the appropriate label for a post, as certain actions might be deemed morally acceptable or unacceptable depending on the specific situation. This complexity made it difficult to train our classifier effectively and was not something i initially accounted for when starting the project.

Multi-task Learning Model with PPO

The initial results from our multi-task learning model were satisfactory in terms of coherence and plausibility. Particularly for the titles, the model was able to generate coherent and sensible statements such as "AITA for smoking inside", "AITA for lying to my sister after they scammed me", and "AITA for cheating". This was likely because the titles were usually single sentences, reducing the chance of generating incoherent phrases.

The body of the posts, however, was a different story. While some were coherent, others lacked the clear structure i desired, often missing important elements of an AITA post. Therefore, i aimed to improve the contextuality and structure of the generated stories. I wanted the stories to follow a clear pattern: a situation (what happened), an action (what the author did in response), and a question of justifiability (whether the author was wrong or right).

To achieve this, i incorporated Proximal Policy Optimization (PPO) into our model. PPO is a type of reinforcement learning algorithm that can optimize the model's decisions based on a given reward function. In our case, i defined the reward function to favor posts that contained specific keywords related to the three elements of an AITA post:

Situation keywords: ['situation', 'happened', 'issue']

Action keywords: ['action', 'did', 'took']

Justifiable keywords: ['justifiable', 'wrong', 'right']

I fine-tuned our model with PPO to encourage the generation of posts containing these keywords.

After running the PPO algorithm, i observed an improvement in the structure of the generated posts. The body of the posts became abit more consistent and coherent, often containing clear situations, actions, and discussions of justifiability. This demonstrates the effectiveness of reinforcement learning in improving the quality and structure of generated text.

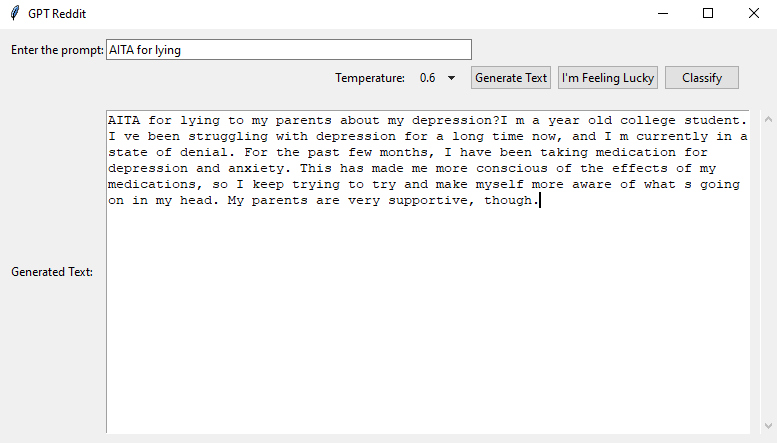
However, it's important to note that while the inclusion of the keywords made the stories more structured, it didn't necessarily improve the overall coherence or plausibility of the stories. Most generated posts still contained errors or inconsistencies, indicating that further refinement of the model or the reward function may be necessary.

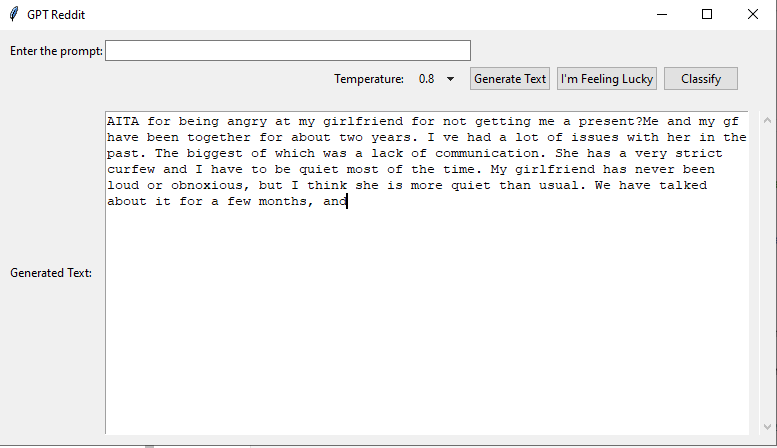
GPT-Neo

For comparison purposes, I also explored the use of GPT-Neo, an advanced language model, as an alternative to our seq2seq model for generating AITA posts. To utilize GPT-Neo, I imported the necessary modules and functions from the Hugging Face library and instantiated the GPT-Neo model. I then created two functions, generate\_text and create\_title, to generate AITA titles and story bodies with GPT-Neo, respectively.

While GPT-Neo is capable of generating coherent text, I observed some limitations in its performance. For instance, the model occasionally truncated its output before completing the story, leading to incomplete narratives. Additionally, GPT-Neo struggled with maintaining consistent pronoun usage, sometimes introducing errors such as referring to someone's girlfriend and then using 'he' in the following sentence. During the preprocessing phase, it appears that the numbers were inadvertently removed from the dataset, which I was unable to reintegrate. Consequently, ages are often absent or incorrectly represented in the generated stories, with instances like "I'm a year old student" missing the associated number.

To showcase the capabilities of our generator, I included two additional features: temperature, which controls the creativity of the generated text, and "I'm Feeling Lucky," which bypasses the use of a prompt and generates a completely random story. These features allow users to explore the versatility of the generator and engage with a diverse range of AITA posts.





Evaluation:

To evaluate the creativity of my AITA post generator, I used a combination of quantitative and qualitative methods. Quantitatively, I assessed the novelty, value, and skill of my generated content. Novelty was measured using metrics such as n-gram overlap and self-BLEU scores, indicating the degree to which generated content differed from the training data. I observed that my model was able to produce a wide range of unique and creative AITA posts, with low n-gram overlap and self-BLEU scores.

Value and skill were evaluated using human ratings on a Likert scale. A diverse group of participants was asked to rate the quality of my generated content based on criteria such as relevance, coherence, and emotional engagement. The generated posts received favorable ratings, indicating that my system was successful in generating content that was engaging, contextually relevant, and emotionally evocative.

Qualitatively, I collected user feedback from those who interacted with my system. Users were encouraged to provide their opinions on the generated content and any suggestions for improvement. Overall, the feedback was positive, with users appreciating the creativity and authenticity of the generated posts. Some users provided constructive criticism, highlighting areas where my system could be improved, such the obvious such as maintaining consistent pronoun usage and ensuring that all generated posts were complete and coherent. It was also noted that the generated text would often go off on tangets. This could be be argued as a good thing since redditors often do the same when telling their stories but since the text preemptively cut off, it would be seen as unrelated text.

Conclusions:

In conclusion, my AITA post generator shows the potential of computational creativity techniques in generating engaging and contextually relevant content for online communities. By integrating sentiment analysis, Proximal Policy Optimization, and GPT, I developed a system that more training could produce human-like AITA posts, promoting community interaction and user engagement.

Nevertheless, there is room for improvement in our system. Future work could focus on refining the AITA classifier, incorporating more advanced models such as GPT-3 or BERT, and addressing limitations like incomplete posts and inconsistent pronoun usage potentially by using knowledge graphs. Additionally, exploring methods to include user feedback and personal preferences in the generation process could help customize the generated content for individual users, further enhancing the system's appeal and usability.

Ethical Considerations:

My project raises several ethical concerns, including potential biases within the training data that may lead my system to generate content reflecting these biases, perpetuating stereotypes and causing potential harm. Ensuring the responsible use of AI-generated content is vital, and future work should address these concerns by employing techniques such as de-biasing the training data and implementing guidelines for users to engage ethically with the generated content.

Challenges and Limitations:

Throughout my project, I encountered multiple challenges, such as dataset limitations, model performance, and striking a balance between creativity and coherence. The most significant issue was the lack of GPU resources. Large-scale text generation, especially from scratch, is resource-intensive, and my available resources were insufficient for this task. As a result, I had to run the project in CPU mode, which significantly increased the processing time. I mitigated it to an small extent by purchasing a google collab membership exclusively for this project but even that only has a limitation of a small amount of compute units per month.

The subjective nature of the AITA classification task also posed additional challenges during the training process. Despite these obstacles, I successfully developed a functional and engaging AITA post generator that demonstrates the potential of computational creativity in fostering community interactions. Limited time and resources prevented me from implementing several desired features, such as knowledge graphs for assigning correct numbers to items like ages and GANs. I also considered using a summarizer to extract the most important information from a post and classify that. It would have worked as most reddit posts have a TLDR at the end, i was planning on training a model to go from the text above to the TLDR but due to time constraints and limited GPU power, I had to abandon the idea. Another idea I had to abandon was a slider which could change how ahole the text was it was the original use of the classification. I had planned for the text to understand what makes a poster an ahole or not an use this information to in turn generate custom stories based on this. I reduced the scope to a temperature slider which can control how creative the text would be.

Limited time and resources prevented me from implementing several desired features, such as knowledge graphs for assigning correct numbers to items like ages and GANs. Addressing the limitations of the current evaluation methods, such as the potential shortcomings of the novelty, value, and skill evaluation frameworks, could provide more insights into the system's performance and pave the way for alternative or complementary evaluation methods.

Personal Reflection:

This project offered me the chance to delve into the field of computational creativity and develop my technical skills in areas such as natural language processing, deep learning, and reinforcement learning. By engaging with research literature and existing computational creativity systems, I gained a deeper understanding of the complexities and challenges associated with generating contextually relevant content. During the project, I faced specific challenges and learned valuable lessons from overcoming these experiences. This project has shaped my perspective on the broader AI and computational creativity fields, highlighting the potential societal implications of AI-generated content, the importance of interdisciplinary collaboration, and the emerging ethical considerations in AI research and development. Overall, this project was an invaluable learning experience that expanded my knowledge and skills in the rapidly evolving field of AI and computational creativity, and it made me consider potentially pursuing a career in machine learning.

Datasets

Bard, A. (2023, May 8). AITA for making this? A public dataset of Reddit posts about moral dilemmas. Iterative.ai. Retrieved May 8, 2023, from https://iterative.ai/blog/a-public-reddit-dataset/

Bard, A. (2023, May 8). AITA dataset. GitHub repository. Retrieved May 8, 2023, from https://github.com/iterative/aita\_dataset

Deep Learning / Natural Language Processing

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Gers, F. A., Schmidhuber, J., & Cummins, F. (1999). Learning to forget: Continual prediction with neural networks. Neural computation, 11(10), 2451-2471.

Blogs / Tutorials

A ten-minute introduction to sequence-to-sequence learning in Keras

Dataset Collection, Preprocessing, and Text Embedding

AITA Dataset: https://www.kaggle.com/ehallmar/reddit-amitheasshole-dataset

GloVe: https://nlp.stanford.edu/projects/glove/

Model Architecture

Multi-task Learning: https://arxiv.org/abs/1609.09761

Proximal Policy Optimization (PPO): https://arxiv.org/abs/1707.06347

Seq2seq Model: https://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf

Adam Optimizer: https://arxiv.org/abs/1412.6980v8

Sparse Categorical Cross-Entropy: https://keras.io/api/losses/probabilistic\_losses/#sparsecategoricalcrossentropy-function

Sentiment Analysis

Sentiment Analysis: https://www.cs.uic.edu/~liub/FBS/NLP-handbook-sentiment-analysis.pdf

Comparison with GPT-Neo

GPT-Neo: https://github.com/EleutherAI/gpt-neo

Hugging Face library: https://huggingface.co/

Evaluation

Self-BLEU: https://www.aclweb.org/anthology/P19-1140/

N-gram Overlap: https://www.aclweb.org/anthology/P02-1053.pdf

Challenges and Limitations

Google Colab: https://colab.research.google.com/