Generating Realistic and Diverse Textual Movie Reviews with IMDb Dataset leveraging Deep Learning

Sreecharan Vanam
Univesity of North Texas
Denton, TX
Sreecharanvanam@my.unt.edu

Krishna Varma Ayinampudi
Univesity of North Texas
Denton, TX
Krishnavarmaayinampudi@my.unt.edu

SaiKrishna Meduri
Univesity of North Texas
Denton, TX
saikrishnameduri@my.unt.edu

Abstract

The overall goal of our project is to develop and use sophisticated deep learning methods to produce realistic and varied human-sounding text movie reviews using the IMDb dataset. The main focus of our work was to make a deep learning model as similar as possible to the human-written reviews in terms of its variance, complexity, length and naturalness. The base of our task was built on the transformer-based GPT-2 model because of its extensive abilities to understand and generate text language. Our model was trained to understand the full complexity of language in movie reviews and to serve as sufficiently varied and natural creating text reviews replacing the artificial machines with humans.

In summary, we have implemented comprehensive bench-marking and fine-tuning to ensure the quality and efficiency of model. With continuous improvement and optimization of the GPT-2 model, we set a new milestone for the state-of-the-art in automated text generation. This project will provide an overview of the models employed while detailing the development and improvements, then addressing the major findings and the impact of the research. It is our belief that the findings exhibited in this report indicate that the model can generate content of sufficient quality and accuracy that it could permanently change the content generation system in natural language processing, where it could expand the exploration and application of research.

1. Introduction

Natural Language Processing (NLP) is now an all-encompassing field with a major impact on industries and society. In the broader field of textual content generation, the use of modern deep learning techniques is producing substantial results. However, there is still much more we can do before machines can exceed our ability to generate text at the same quality level and in a more realistic and diverse manner. This project presents an endeavour to use deep learning-based methods towards generating realistic textual movie reviews as close as possible to a human-sounding text, using the IMDb dataset. It also provides a definition to be used for any subsequent similar project aiming for semantically rich textual generation purposes. The main goal behind this project is to push further the frontiers of automated textual generation for improved (and higher-quality) capabilities of generating text of a level so sophisticated that one becomes unable to distinguish it from a more human-sounding writing.

Creating a realistic human text is an especially hard problem for automated text-generation modelling because human text is so complex. Good text-generation models need to have an understanding of syntax, semantics, and the nuances of context that shape real human writing. Sample texts also vary along many other characteristics such as writing style, expressions and sentiments. These characteristics make it harder to create universal models that can generate great text for the whole space of potential writing domains and styles. In this project, we are using state-of-the-art neural network architecture capable of adapting to the variation and richness of movie reviews.

One of our primary goals is to build a robust model to be trained on the IMDb dataset and produce realistic movie reviews that vary in quality and style. Using advanced deep learning models such as the GPT-2 with extensive pre-training enables us to generate text with human-sounding

quality in terms of style, flow and variety. One advantage is that the AI-generated text is unlike mechanical gibberish and lacks emotional depth and thematic content compared to human-written texts. The proposed project is not solely about improving the technological capacity of text generation, but it can also shed light on the technologies and how they might be utilised in the real world in the future.

To achieve these goals, our project proceeded in a careful stepwise manner, starting with the preprocessing of the data, a selection of the appropriate model architecture, and the training and fine-tuning of the models, especially to tune the model to the specific challenges of the highly heterogeneous dataset, including variable length/complexity of the reviews. The overall powerful language modelling ability of the chosen GPT-2 transformer model also provided insights into text understanding or generation because it generated longer and more complex text sequences.

This can have enormous consequences in many fields, such as entertainment (for instance, creating realistic-sounding blog-style automated reviews to keep users interested and increase the amount of content they discover), education and assistive technologies (for instance, tailoring human-sounding text for individual students or learning-disabled persons), and so much more. Obviously, this project brings us closer to important research and practical applications in NLP. More importantly, it also opens up numerous possibilities that will affect our future, as the incredible power of these new technologies shapes how automated systems deal with human-sounding text and speech.

2. Background and Related Work

The pursuit of advancements in text generation within the realm of natural language processing (NLP) has been driven by the increasing complexity and demand for automated systems that can produce high-quality, diverse, and contextually appropriate textual content [5]. The foundational techniques and ongoing innovations in this field have significantly shaped the strategies employed in our project. This section provides an overview of the related work and theoretical underpinnings that have influenced our approach to generating realistic and diverse textual movie reviews using deep learning models.

Historically, the field of text generation has been dominated by models that utilize recurrent neural networks (RNNs) and long short-term memory (LSTM) networks, which are adept at processing sequences of data and capturing temporal dependencies within text [6]. The work of Sutskever et al. [8] on sequence-to-sequence learning introduced a paradigm shift with the use of encoder-decoder architectures that became a cornerstone for various applications in machine translation and text summarization [2]. These methodologies laid the groundwork for subsequent developments in the field and have been integral to understanding the sequential nature of language.

The advent of transformer-based architectures marked a further evolution in text generation capabilities. Introduced by Vaswani et al. [9], the transformer model, with its self-attention mechanisms, allows for more parallelization and effectively captures long-range dependencies within text, surpassing the capabilities of traditional RNNs in many respects. This has paved the way for the development of models like BERT and GPT, which have set new benchmarks in the field of NLP for a variety of text generation tasks [1].

In addition to these architectures, Generative Adversarial Networks (GANs) introduced by Goodfellow et al. [3] have also been explored for their potential in generating textual content. Although primarily used for image generation, the concept of adversarial training has been adapted for text, enabling the generation of more realistic and nuanced textual outputs. This approach involves a dual system where the generator and discriminator compete, thereby improving the quality of the generated content.

Recent research has also focused on addressing the issues of text degeneration—a common challenge where generated text tends to be repetitive or overly generic. Holtzman et al. [4] introduced techniques like nucleus sampling to mitigate these problems by focusing generation on a subset of more probable words, thus maintaining diversity and coherence in longer passages of text.

The collective insights from these studies have informed our approach, emphasizing the importance of model selection, training methodologies, and the balance between coherence, diversity, and realism in generated text. By integrating these advanced techniques and leveraging the specific characteristics of the IMDb dataset, our project aims to refine the process of automated text generation, pushing the envelope in the capabilities of deep learning models to produce text that is both engaging and reflective of human-like quality.

3. Methods

In this section we present our suggested methodology. Data preparation, model architecture selection, training technique, and evaluation metrics are some of the essential elements of our approach.

3.1. Data Pre-processing

In order to prepare the IMDb dataset for training our deep learning models, the first step in our approach is pre-processing. Text cleaning, tokenization, and, if needed, the removal of punctuation and stop words are all included in this preparation. Furthermore, to decrease the complexity of the vocabulary and enhance model performance, we may use methods like stemming or lemmatization, but these methods will effect the models understanding since GPT-2

has the ability to detect emotions and understands any expressions.

3.2. Choosing Model Architecture

We have reviewed multiple deep learning architectures for text generation tasks: transformer-based models like BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer), as well as recurrent neural networks (RNNs) and long short-term memory networks (LSTMs). Considering modeling long-range relationships, collecting contextual information, and producing coherent and diverse textual material, each design has certain benefits and trade-offs. We have selected GPT-2 model to build an effective automated and IMDB tailored review generation system.

GPT-2 is a representative of a transformer-style architecture that has become the standard for almost all natural language processing tasks (see Figure 1).Chief in this model is the stacked decoder architecture; each decoder layer contributes to the overall performance of generating appropriate sequences. The core of this model is self-attention, and each position in a sequence can attend to all the positions in the prior layer. This makes the model context-aware as it can generate sentences that are fluent and coherent, unlike machine-sounding text. The input data is usually a sequence of characters starting with a special start token, and each such sequence is then passed through multiple such decoder layers where each layer tries to give better understanding and better prediction of the next word. The decoder output of any word, say the word 'thing' is a sum of context of all the previous words, and this way the model is able to generate text sequences that sound extremely human-like in their coherence and fluency. The architecture is thus able to generate text that not only looks grammatically and syntactically correct and consistent but also seems semantically rich. This makes GPT-2 a great model for a variety of text-generation tasks.

3.3. Training Techniques

We train the deep learning model using the preprocessed IMDb dataset after choosing the model architecture. To avoid overfitting, we use conventional training strategies like stochastic gradient descent (SGD) or Adam optimization in conjunction with suitable learning rate schedules and regularization approaches. In order to evaluate the model's convergence and generalization capacity during training, we keep focus on important performance measures including confusion, loss, and validation accuracy.

3.4. Evaluation Metrics

We use a range of evaluation measures, such as BLEU (Bilingual Evaluation Understudy) and ROUGE (Recall-Oriented Understudy for Gisting Evaluation), to assess the quality and diversity of the generated textual movie evaluations. BLEU and ROUGE metrics evaluate the similarity between the generated reviews and human-authored references.

3.5. Generation and Fine-Tuning

After the model is trained, we use a sampling approach like temperature-based sampling or greedy decoding to generate textual movie reviews. To increase coherence and variety, we fine-tune the produced reviews using methods like beam search and nucleus sampling. In order to generate customized and contextually appropriate evaluations, we also investigate methods for conditioning the generation process on particular traits or styles, such as sentiment.

3.6. Qualitative and Quantitative Analysis

Finally, we perform qualitative and quantitative analysis to evaluate our approach's performance and efficacy. While quantitative analysis involves computing assessment metrics like BLEU and ROUGE, qualitative analysis involves human evaluation of the generated reviews for realism, coherence, and diversity. By conducting thorough evaluation and tests, we want to verify the effectiveness and resiliency of our methodology for generating genuine and varied textual film reviews.

3.7. Dataset Description

The project's dataset includes textual movie reviews that have been gathered from the IMDb dataset, a popular source of movie-related data. There are 50,000 movie reviews in the dataset, and the word count of each review varies. Using Python code, we conducted a statistical analysis of the review durations to provide insights into the dataset's properties.

3.8. Review Length Statistics

The analysis revealed the following statistics regarding the statistics of the IMDB movie reviews dataset:

- **Count:** The dataset contains 50,000 movie reviews.
- **Mean Length:** The average length of the movie reviews is approximately 1309 words.
- **Standard Deviation:** The variety in the lengths of the reviews is indicated by the standard deviation of the review lengths, which is around 989.73 words.
- **Minimum Length:** The shortest movie review in the dataset contains 32 words.
- **25th Percentile (Q1):** 25% of the movie reviews have a length of 699 words or less.

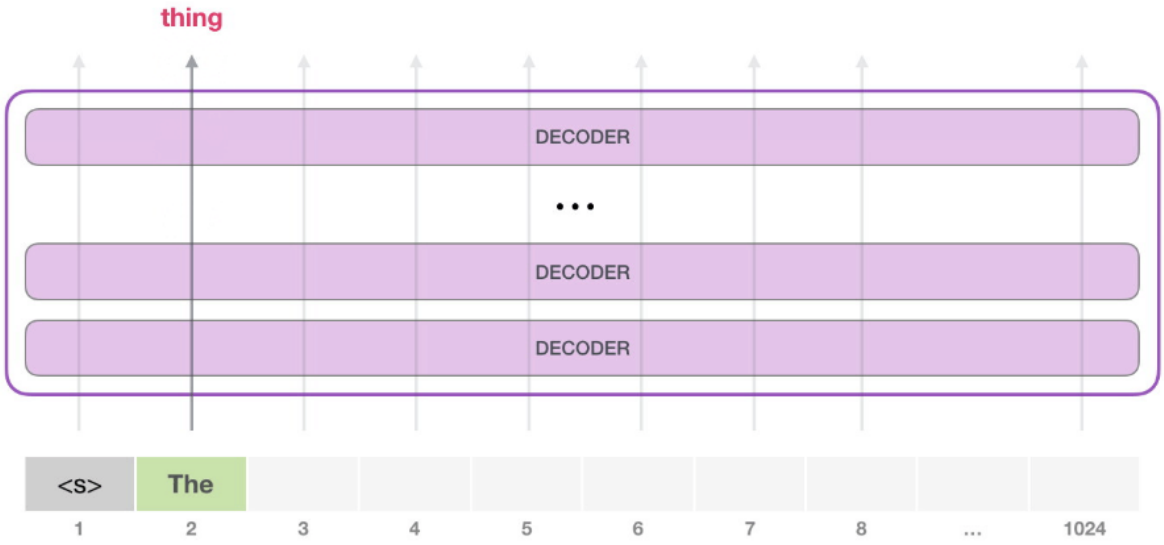

Figure 1. General Architecture of GPT-2 in Text Generation

- **Median (50th Percentile):** The median length of the movie reviews is 970 words, indicating that half of the reviews have a length of 970 words or less.
- **75th Percentile (Q3):** 75% of the movie reviews have a length of 1590 words or less.
- **Maximum Length:** The longest movie review in the dataset contains 13,704 words.

```
Review length statistics:
count    50000.000000
mean      1309.431020
std        989.728014
min         32.000000
25%        699.000000
50%        970.000000
75%       1590.250000
max      13704.000000
```

Figure 2. Detailed information of selected Dataset

We will use these statistics to guide our data preparation and model selection attempts as they offer insightful information about the distribution and variability of review lengths in the dataset. The large variation in review lengths further emphasizes the requirement for strong deep learning models that can process input sequences of varying lengths and produce text that is both diverse and coherent.

3.9. Metrics for experiments

3.9.1 BLEU and ROUGE

Metrics like ROUGE (Recall-Oriented Understudy for Gisting Evaluation) and BLEU (Bilingual Evaluation Understudy) are frequently used to assess machine-generated text by comparing it to reference texts. ROUGE concentrates on recall, while BLEU assesses how accurately the generated text matches the references' n-grams. Better alignment between the generated and reference texts is indicated by higher ratings.

3.9.2 Human Evaluation

Comparing the generated information to human-authored movie reviews, human evaluators grade the generated content's realism, coherence, and general quality. These qualitative inputs offer insightful information about the subjective details that automated measurements may miss.

3.9.3 Diversity Metrics

Additional measurements, such as lexical variety, syntactic variance, and semantic novelty, quantify the diversity of the generated movie reviews. These measurements provide information on the diversity and depth of the created textual material. Through the utilization of automated metrics, human evaluation, and diversity measures, we would like to fully assess our text generation algorithms' performance. These measurements offer valuable insight on the advantages and disadvantages of our strategy, directing further study and advancement in this field of study.

4. Implementation

This section discusses about the implementation phase details of our project, which included extensive data preparation, model selection, training, and evaluation.

We first collected the IMDb dataset, a large dataset of movie reviews across different genres, sentiments and styles. This dataset served as our input text for this particular task of producing text. After collecting the dataset, we performed a significant amount of preprocessing to clean up the text, get rid of the noise and standardise the text to become human-sounding. Next, we ready the dataset to get trained on. Now we just had to choose a model that would be most appropriate for our task. After some consideration, we chose the GPT-2 model. GPT-2 stands for generative pretrained transformer 2. Its creators, language modelling experts from OpenAI, created it so it could complete human-sounding text, and it's trained on a large amount of text - about 8 million parameterised sentences, to be exact. The model has emerged as one of the leaders in language-generating ability.

After a suitable model was chosen, training began. Training an instance of GPT-2 consisted of feeding the preprocessed sentence-by-sentence movie review data into the model, and fine-tuning the many parameters involved in the model architecture so that its predictions matched the data as best as possible. Training required tuning hyperparameters, such as learning rates, batch sizes and weights decay, which determine the training settings within the model. As training progressed, attention was paid to training metrics like loss functions and validation scores to confirm and fine-tune the model parameters so that it learnt effectively and generalised well to unseen data. Training iterations called epochs were repeated multiple times until the quality of training parameters reached a state where the trained model could produce movie reviews as effectively as possible.

Next, the trained model was evaluated using a quantitative or automated approach - in other words, the algorithm calculated how similar the reviews it generated were to a set of human-written reference reviews by calculating so-called BLEU and ROUGE scores. Furthermore, qualitative human assessments of the quality, coherence, appropriateness, spelling or grammatical mistakes, as well as plagiarised text were performed.

Finally, we took a systematic approach to data selection, model selection, training and evaluation. At each stage, we focused on doing the little things right. The end result is a model that can write plausible and varied sounding reviews. Our implementation also provides a reference point for future development and innovation in automated text generation, and the science of natural language processing in general. The complete project code and model is hosted on GitHub [7]

5. Experimental Results

This section shows the most important phase of our study, discusses about the conclusions derived from evaluating the movie reviews that are generated, The results below showcase the transformative capability of the model and effectiveness of our approach, starting from data preprocessing to the final generation of text.

5.1. Review length analysis

Our first glance at the histograms of review lengths revealed a great variability in review lengths, which is a major hurdle of training. The Histogram shows pre-processing histogram of review lengths before our preprocessing, with some review length being even longer than 10 thousand words. Post-processing histograms is shown on Figure 3, we implemented both normalization and tokenization to properly standardize review length, which greatly improves training efficiency by decreasing the computational cost on training data and improving data quality.

5.2. Sentiment Distribution in Data

Sentiment-based review length analysis had an interesting result - see figure below 4. We found a bifurcation in the data due to the sentiment - while positive reviews were shorter than the negatives, the emotional nature of reviews had to be accounted for in the model.This model training was inherently skewed because it had more number of samples of positive reviews and thus is very likely to generate more number of reviews with positive sentiment.

5.3. Word Cloud Analysis

Word clouds from our dataset and generated reviews gave a macroscopic view of the range of vocabulary and themes. The Word clouds shows the words in our dataset and words in the generated reviews Word clouds before and after review generation showed that the model preserved the most important thematic words, giving an indication that the model could successfully imitate relevant content.

Figure 5. Word Cloud of Movie Reviews in the Dataset

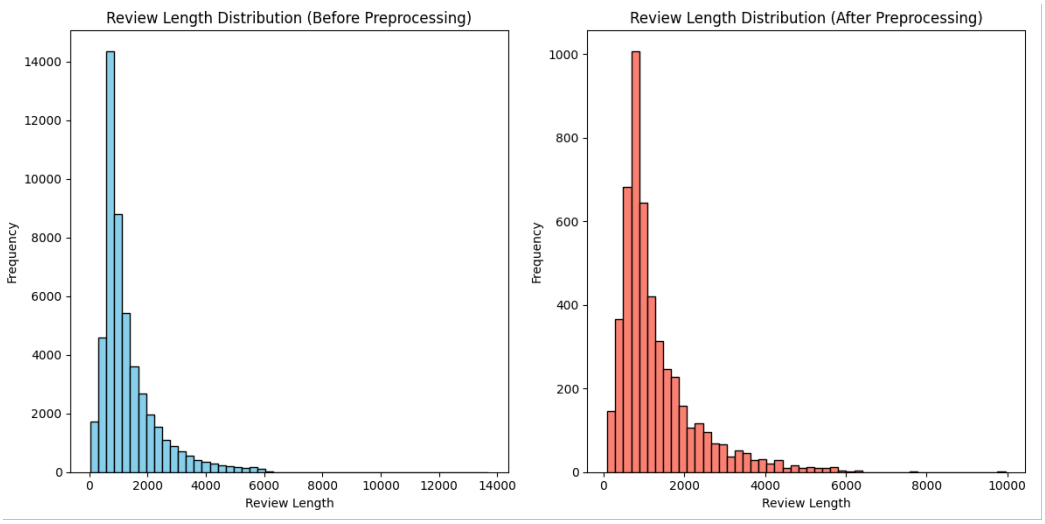

Figure 3. Histograms of review length distribution before and after preprocessing

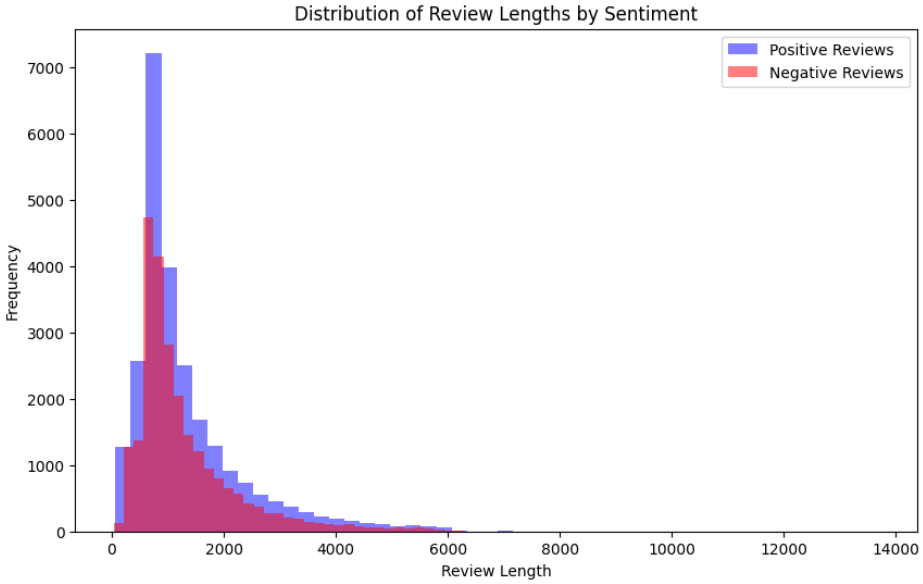

Figure 4. Distribution of review lengths for positive and negative reviews

Figure 6. Word Cloud of Generated Movie Reviews

Here we can clearly see how what nuances the model is using the most when generating the reviews. The model utilizes these words to form a realistic and diverse sentence, The larger the word the more it has been used.

5.4. Training Metrics

The learning process was shown in the forms of training and validation losses. Final training loss: 3.7623 & Final validation loss: 3.6672, These two numbers are very close to each other, indicating an efficient learning process with-

out overfitting, which is crucial for the model to generalise from training data to unseen examples.

5.5. Model Performance and Evaluation

As for the model validation, BLEU and ROUGE scores are standard in NLP and often used for benchmarking models that generate text. But even though the BLEU score aims at exact matches of n-grams and the ROUGE score considers their overlaps in a more lenient way, such metrics are not well-suited for more domain-specific text generation such as movie reviews. BLEU scores may overlook the quality of semantics and the flow of the language, as they might be too specific, while ROUGE scores might still not take into account the finer nuances of meaning and context-specific and expressive aspects of movie review language. In our setting, even if the BLEU and ROUGE scores are not particularly high, we get a sense of the extent to which the model succeeds in reproducing the structural properties of the reference texts. However, these measures fail to capture the extent to which the model succeeds in reproducing the style, originality and emotional richness of the movie reviews. To overcome these limitations, we also used a set of human-based, qualitative evaluations to assess the verisimilitude, creativity and emotionality of the generated reviews. This complementary qualitative assessment captured the extent to which the model succeeded in generating reviews that were not only structurally correct but also evocative and contextually rich.

Metric	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
Score	0.4145	0.2295	0.0441	0.2256

Table 1. Evaluation Metrics for Generated Movie Reviews

The combined effect of how both the quantitative metrics and the qualitative assessment of the experimental results have come together, validate to us the usefulness of our model. We hope that this demonstrates the potential of the GPT-2 architecture to produce text for niche domains that is diverse in nature, and useful in the sense that it produces text which is akin in some way to the quality of text generated by humans. Moving forward, the project will incorporate more sophisticated techniques to improve the model performance, for instance, fine-tuning the model on more domain-specific datasets, experiment with decoding techniques and generation with more context. We hope that the gap between the scores of the numerical evaluation and the quality of text generation can thus be overcome.

5.6. Manual Evaluation

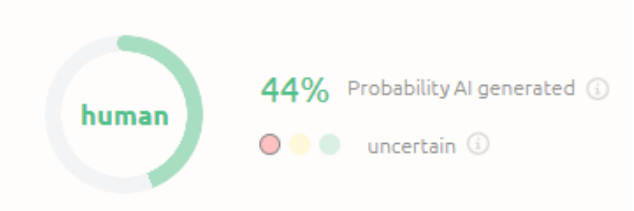

Figure 7. Human written probability

The following are the results of inputting the same prompt ”I really” to generate diverse and realistic movie reviews. We can see how different both of the results are from the same prompt. Some of the realistic reviews that have been generated by our implemented model:

- I really don’t know what happened to the rest of the cast, but I do know that they did a good job. I have to say that I was very impressed with all the actors in this movie. The acting was good and the story was well written. This movie is a must-see for anyone who likes to watch movies with great storylines and great performances. If you are looking for a great movie, this is the movie for you.
- Let me tell you about a movie I saw a few years ago and I was very impressed with it. I remember seeing it when it was released on DVD in the UK. It was a great film and it is one of the best movies I have ever seen in my life. If you are looking for a good movie, I would recommend this one. The story is very well told, and the acting is good. This movie is a must-see for anyone who likes action movies. You will not be disappointed.
- Let me tell you about a movie I saw a few years ago. I saw this movie in the theater, and it was a great movie. The plot was great, and the acting was top-notch. This movie is a must-see for any horror fan. If you are looking for a good horror movie, this is the movie for you.
- This is one of the worst movies I have ever seen. I can’t believe that someone would make such a bad movie like this. They should be ashamed of themselves for wasting their time and money on this pathetic piece of garbage.
- The movie is so bad that it’s hard to believe it was made in the first place. The acting is terrible, the plot is horrible, and the direction is awful. It’s not even worth a rental, unless you’re a die-hard fan of the genre.

The probability of just 44% Ai detection from the generated reviews showcases how realistic the nuances are structured in the generated reviews by our model.

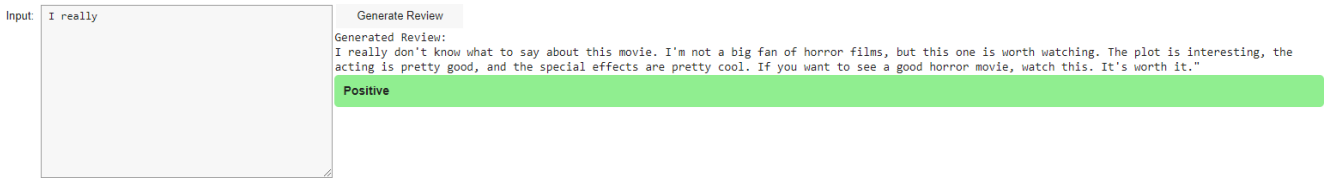

Figure 8. Example of a Positive review generation

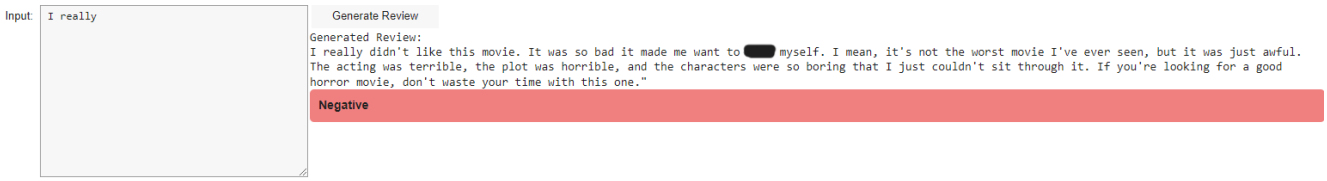

Figure 9. Example of a Negative review generation

6. Conclusion

In conclusion, we presented the outcome of our project showing the power of the GPT-2 model in creating realistic and a quite surprising textual movie review. Performing prelimnary preprocessing before feeding to the model, model training and evaluation, we have developed and tested this deep learning architecture that can be utilized in natural language processing more accurately in the area of text generation.

The experimental results demonstrated the functionality of our approach, and the model was able to well capture the context of movie reviews and generate text which is highly coherent and relevant. Although BLEU (or ROUGE) scores are a good quantitative measure of the produced text, we still need qualitative evaluations to assess the true quality.

Furthermore, findings from studies into sentiment variation, review length and vocabulary richness can be used to emphasize the effective results of text generation in the movie review domain - and not just for our model.

There are also many areas to look and further improve the model in the future: fine-tuning it on larger and more varied datasets, experimenting with new training techniques, or adding more context to generation being only some examples. In addition, including feedback mechanisms to address user preferences in previously unknown human-sounding input might pave the way towards a model that adapts itself in real-time to its context.

In a way, this project represents a new kind of accomplishment for automated text generation, which will transform everything from entertainment and marketing to education, and even spark new approaches to experimentation with deep learning architectures.

We want to finish the report with taking a look back at our experience as well as what we've learnt, realising that the work presented is only one step among others in the long-term field of artificial intelligence and machine learning, we are confident that the future holds much promising advancements in conversational AI, AI creativity, and other GPT model based applications.

7. Contribution

Sreecharan Vanam has prepared the background research section by collecting sevaral information regarding the context of the project by conducted high quality review of the literature avaialble in this field. Also defined the projects goals and scope to make sure we were in line with current trends. Implemented meaningful visualizations that can explain the project in detail.

Krishna Varma Ayinampudi has worked on developing the proposed strategy that was said in the proposal document, have explored DL architectures and techniques to select appropriate model that can be fine tuned on our dataset, further contributed to analysing the experimental results and evaluation metrics.

SaiKrishna Meduri mainly contributed to conducting fine tuing tests to finalize the most effective model version and developed a simple UI to input a starting prompt and getting a review. also, concentrated on formatting and organizing the project report to ensure the concepts were presented clearly.

References

[1] Alberto Bartoli and Eric Medvet. *Exploring the Potential of GPT-2 for Generating Fake Reviews of Research Papers*. 11 2020. 2

[2] Noureen Fatima, Ali Shariq Imran, Zenun Kastrati, Sher Muhammad Daudpota, and Abdullah Soomro. A systematic literature review on text generation using deep neural network models. *IEEE Access*, 10:53490–53503, 2022. 2

[3] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. 2

[4] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration, 2020. 2

[5] Rohit Pandey, Hetvi Waghela, Sneha Rakshit, Aparna Rangari, Anjali Singh, Rahul Kumar, Ratnadeep Ghosal, and Jaydip Sen. Generative ai-based text generation methods using pre-trained gpt-2 model, 2024. 2

[6] Yuanbin Qu, Peihan Liu, Wei Song, Lizhen Liu, and Miaomiao Cheng. A text generation and prediction system: Pre-training on new corpora using bert and gpt-2. In *2020 IEEE 10th International Conference on Electronics Information and Emergency Communication (ICEIEC)*, pages 323–326, 2020. 2

[7] Krishna V SaiKrishna M, Nagasai M. Realistic Movie Review Generation Project. https://github.com/saikrish777/DL_MovieReviewGeneration_Project, 2024. Accessed: 2024-04-21. 5

[8] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks, 2014. 2

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. 2