Movie Recommender:

Utilizing BERT4Rec

**Group Name:** Bill Greats

**Team Members**: Kumar Kishalaya, Sakshi Arya, Shiv Goel, Ruben Casas

Table of Contents

[**Executive Summary 3**](#_z3kr48zhu7i8)

[**Introduction 4**](#_4gq3g0k1c7u5)

[**Business Objective 5**](#_sqday0njiudv)

[**Key Actionable Business Objectives 5**](#_1k2b833mz3yt)

[**Our Clients and Services 6**](#_k99hy17a982r)

[**Metrics of Success 7**](#_t37gzcgxcim0)

[**Role of Analytics 7**](#_iqab1v65lnyj)

[**Executing the Analytics 7**](#_b2xuftxztfzk)

[**Results 11**](#_qo7uv5twbdug)

[**Future Scope 13**](#_rqcqzckn6uy8)

[**Limitations 14**](#_mex2uxhske5f)

# 

# 

# 

# 

# 

# 

# 

# 

# 

# 

# **Executive Summary**

In the ever-evolving landscape of recommendation systems, accurately modeling users' dynamic preferences from their historical behaviors poses a significant challenge and is crucial for enhancing user satisfaction and engagement. Recommendation systems have become ubiquitous, permeating various aspects of our daily lives, from suggesting videos on YouTube and recommending products on Amazon to curating playlists on Spotify and suggesting movies on Netflix. Despite their widespread presence, the core methodologies underpinning these systems have seen relatively slow progress. Traditional methods, which employ sequential neural networks to encode users' interactions from left to right into hidden representations, often fall short due to their unidirectional nature and rigid assumptions about sequence order. These limitations restrict the power of hidden representations in users' behavior sequences and assume a rigidly ordered sequence that is not always practical in real-world scenarios.

Given the central role of recommendation systems in driving user engagement and satisfaction, there is a pressing need for innovation and experimentation in this field. Traditional models are often constrained by their inability to fully capture the complex, bidirectional relationships inherent in user interactions. This results in suboptimal recommendations that can negatively impact user experience and engagement metrics. To address these limitations, we introduce BERT4Rec, a sequential recommendation model that leverages deep bidirectional self-attention mechanisms. By adopting the Cloze objective for training, BERT4Rec predicts random masked items within user behavior sequences, utilizing context from both the left and right sides of each item. This bidirectional approach allows for a more comprehensive understanding of user behavior, capturing dependencies that unidirectional models miss.

This report outlines the methodology, implementation, and performance of BERT4Rec on the MovieLens dataset, demonstrating its superiority over state-of-the-art models through extensive experiments. The findings highlight BERT4Rec's ability to significantly enhance recommendation accuracy, thereby offering a promising solution for personalized content delivery. By improving the way recommendation systems understand and predict user preferences, BERT4Rec can lead to more relevant and engaging content suggestions, ultimately driving higher user satisfaction and loyalty.

Furthermore, the need for continuous experimentation and innovation in recommendation systems cannot be overstated. As user expectations evolve and the volume of data grows exponentially, recommendation systems must become more sophisticated to meet these demands. By exploring new models like BERT4Rec and pushing the boundaries of current methodologies, we can develop more robust and effective recommendation systems that cater to the nuanced preferences of users. This drive for innovation is essential for maintaining a competitive edge in the digital landscape, where user experience is a key differentiator.

# **Introduction**

In today’s digital age, recommendation systems have become the backbone of many online services, seamlessly integrating into our daily lives. From suggesting videos on YouTube to recommending products on Amazon, and curating personalized playlists on Spotify, recommendation systems are everywhere. They are designed to enhance user experience by providing tailored content that aligns with individual preferences. This personalization drives user engagement and satisfaction, which are critical for retaining users and boosting revenues. For instance, Netflix, which is renowned for its recommendation system, credits its algorithm for saving the company approximately $1 billion annually by reducing churn and keeping users engaged.

The market for recommendation systems is expanding rapidly, fueled by the growth of e-commerce and streaming services. According to recent studies, the global recommendation engine market size is expected to reach $12.03 billion by 2026, growing at a compound annual growth rate (CAGR) of 37.46% from 2019 to 2026. This explosive growth underscores the importance of effective recommendation systems in driving business success.

Despite the critical role these systems play, the underlying technology has seen relatively slow progress. Traditional models often struggle with capturing the complexity of user interactions and the bidirectional nature of user behavior sequences. This has prompted major tech companies to invest heavily in next-generation recommendation systems. For example, Instagram employs cutting-edge machine learning algorithms and utilizes thousands of GPUs to process vast amounts of data, delivering highly personalized content to its users. Similarly, Facebook has developed a sophisticated recommendation system that relies on deep learning techniques to predict user preferences and enhance the user experience.

The current state-of-the-art models, while effective, often fall short in capturing the nuances of user interactions. These models typically use unidirectional sequential neural networks, which encode user interactions in a fixed order, limiting their ability to fully understand user preferences. To address these challenges, we introduce BERT4Rec, a sequential recommendation model inspired by the BERT (Bidirectional Encoder Representations from Transformers) architecture used in natural language processing (NLP). BERT4Rec leverages deep bidirectional self-attention mechanisms to model user behavior sequences, offering a more nuanced and comprehensive understanding of user interactions.

BERT4Rec is designed to overcome the limitations of traditional models by utilizing the Cloze objective for training. This involves predicting random masked items within user behavior sequences by jointly conditioning on their left and right context. By doing so, BERT4Rec can capture complex dependencies and provide more accurate and relevant recommendations. In this report, we apply BERT4Rec to movie recommendations using the MovieLens dataset, which includes interactions between 6,541 users, 3,423 movies, and 100k ratings. This dataset allows us to train and evaluate the model, demonstrating its potential to significantly enhance the accuracy of recommendations.

The need for innovation and experimentation in recommendation systems is more pressing than ever. As user expectations evolve and the volume of data grows exponentially, developing more sophisticated models like BERT4Rec is essential to meet these demands. By pushing the boundaries of current methodologies, we aim to create more effective recommendation systems that cater to the nuanced preferences of users, ultimately driving higher user satisfaction and loyalty. This report outlines the methodology, implementation, and performance of BERT4Rec, providing a comprehensive analysis of its capabilities and potential impact on the recommendation system landscape

# **Business Objective**

The primary goal of this project is to develop a robust and accurate recommendation system that can effectively predict user preferences based on historical data. By leveraging the BERT4Rec model, we aim to enhance the accuracy of movie recommendations on platforms like Netflix and Amazon Prime Video. This system is designed to not only improve user satisfaction but also to increase engagement and retention rates, ultimately driving revenue growth for these platforms. Our objective is to demonstrate that BERT4Rec, with its bidirectional self-attention mechanism, can outperform traditional unidirectional models in capturing the nuances of user behavior. This will be achieved by training the model on a large dataset of user interactions, optimizing its parameters, and evaluating its performance using relevant metrics. By doing so, we aim to provide a compelling case for the adoption of BERT4Rec in real-world recommendation systems.

# **Key Actionable Business Objectives**

To achieve our business objective, we have undertaken several strategic initiatives that harness the power of advanced analytics and innovative technologies:

1. **Developing BERT4Rec**: We have implemented a sequential recommendation model based on the Transformer architecture, which captures bidirectional dependencies in user behavior sequences. This involves designing and training the model to predict masked items in a sequence, leveraging both left and right context to enhance prediction accuracy.
2. **Training and Evaluation**: Utilizing the MovieLens-25m dataset, we train and evaluate the BERT4Rec model, benchmarking its performance against the existing state-of-the-art model. This includes pre-processing the data, constructing user interaction sequences, and applying appropriate evaluation metrics to assess the model's effectiveness.
3. **Optimizing Model Performance**: We focus on fine-tuning the model's parameters and training procedures to enhance its performance. This includes experimenting with different training strategies, such as using a custom vocabulary and training BERT from scratch, to achieve optimal results.
4. **Real-world Application Scenarios**: To demonstrate the practical applicability of BERT4Rec, we provide examples of its use in various scenarios, including movie recommendations in genres like Adventure/Fantasy, Action/Adventure, and Comedy. These examples showcase the model's ability to provide relevant and personalized suggestions based on user history

# **Our Clients and Services**

We collaborate closely with various stakeholders to deliver advanced recommendation systems that enhance user experience and drive business success. Our primary clients and the services we offer include:

**Streaming Platforms**

**Clients**: Services like Netflix, Hulu, and Disney+.

**Services**: Development, deployment, and maintenance of recommendation systems to personalize content delivery, improve user engagement, and increase retention rates. We also provide consulting services to optimize recommendation strategies tailored to their specific needs.

**E-commerce Websites**

**Clients:** Platforms such as Amazon, eBay, and Alibaba.

**Services:** Implementation of advanced recommendation algorithms to suggest products, enhancing the shopping experience and driving sales. Our consulting services focus on integrating BERT4Rec into existing infrastructures and optimizing recommendation strategies to increase conversion rates.

**Content Providers**

**Clients:** Companies offering digital content, such as news sites, blogs, and digital libraries.

**Services:** Customizing recommendation systems to deliver personalized content to users, increasing engagement and satisfaction. We assist in data preprocessing, model training and evaluation, and parameter optimization to ensure the highest accuracy and relevance of recommendations.

**Music Streaming Services**

**Clients:** Platforms like Spotify, Apple Music, and Tidal.

**Services:** Developing personalized playlist curation and music recommendation systems to enhance user experience. Our services include continuous monitoring and optimization to adapt to evolving user preferences and ensure sustained engagement.

# **Metrics of Success**

The degree of success in our initiative will be measured by metrics such as **CTR, click-through rate, and user watch time**. We can measure the click-through rate of movies the recommendation system has presented to the user following the completion of a movie or the recommendation subsection on the home screen. User watch time is the second metric that we aim to measure, keeping our users engaged in movies they want to watch is essential for movie-sharing platforms. Users want to feel they get their value out of platforms by time spent watching. Since this is an experimentation with a new architecture it is difficult to say how much it will help in revenue but it is worth exploring.

# **Role of Analytics**

The role analytics plays in our initiative is measuring feedback on engagement metrics influenced by personalized recommendations, validating the effectiveness of BERT4Rec. The statistical method approach used is A/B testing. This is a go-to statistical method of evaluation that can be implemented to aid in determining the success of utilizing BERT4Rec through our key metrics, CLT and watch time. We will be able to evaluate the success of implementing the BERT4Rec through causal inference and determine the next steps in implementation.

# **Executing the Analytics**

Those responsible for implementing the recommender system will be the data science team and IT department. The data science team will be responsible for training and implementation where the IT department will be responsible for integration and deployment. Regular meetings will be necessary to align metrics definitions and refine analytical approaches based on ongoing results.

**Data Description**

The dataset used in this project is the MovieLens-25m dataset, comprising interactions between 162,541 users and 62,423 movies. This dataset provides a rich source of information, enabling the construction of detailed user interaction sequences. Each sequence represents the chronological order of movies watched by a user, serving as the basis for training our recommendation model. The dataset includes various attributes such as user IDs, movie IDs, timestamps, and ratings, which are used to build the user behavior sequences. By leveraging this comprehensive dataset, we can effectively train and evaluate the BERT4Rec model to provide accurate and personalized movie recommendations

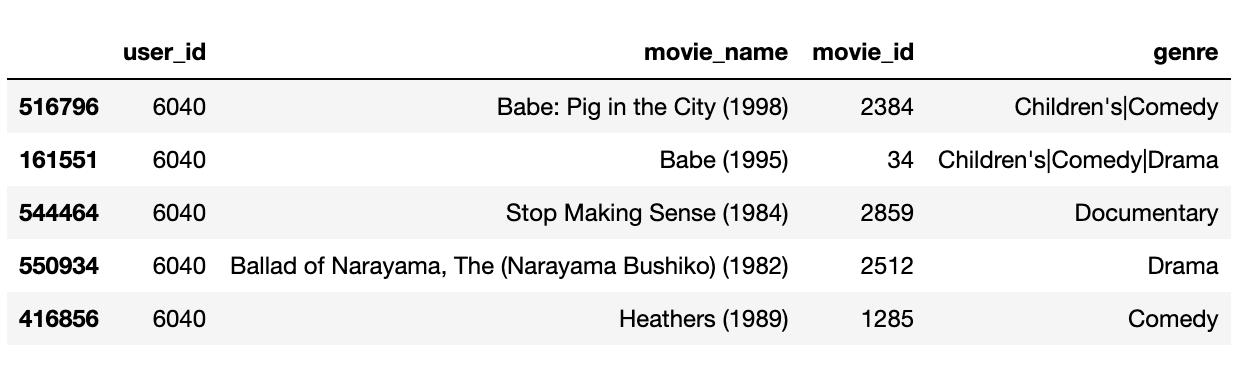
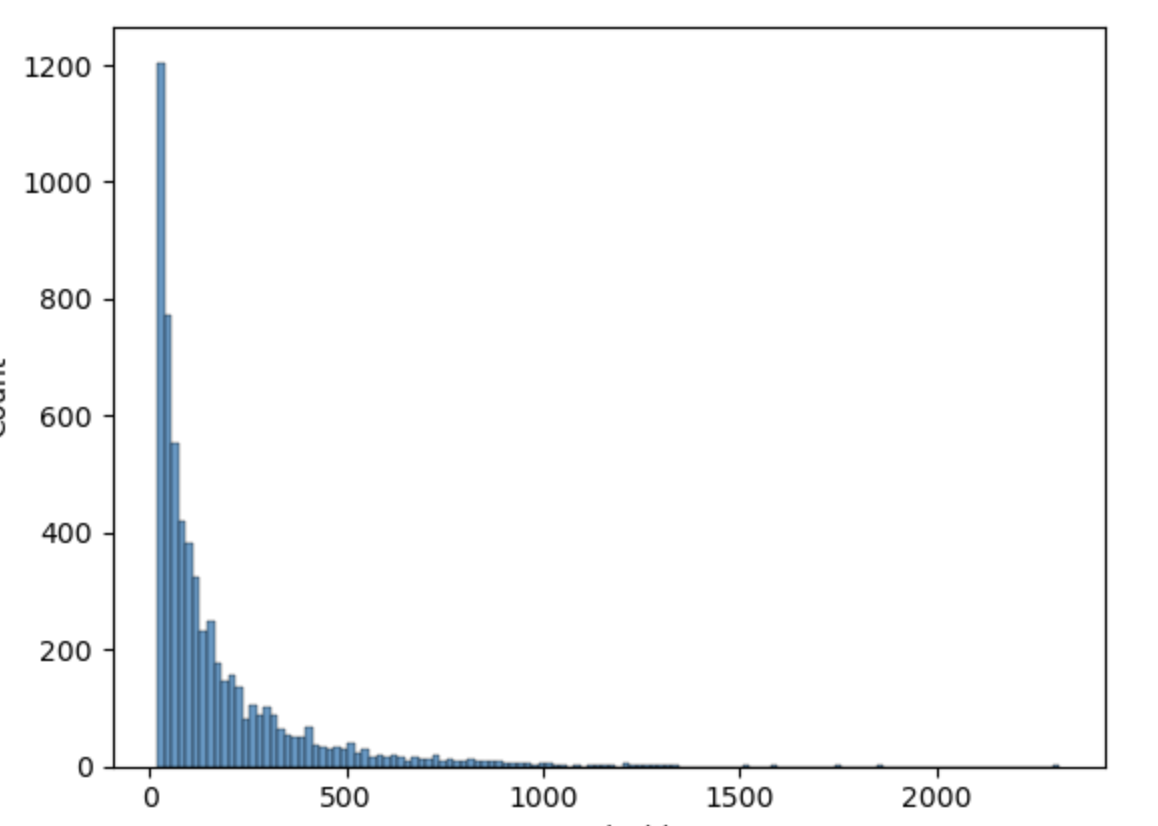


Figure 1- How our dataset looks like



*Figure 2- Most users have under 500 movie watch counts. This will decide our truncation length.*

**Analytics Methodology**

Our analytics methodology involves several key steps that ensure the development of an effective recommendation system:

**Data Preprocessing**: This step involves cleaning and structuring the MovieLens-25m dataset to create time-sorted sequences of user interactions. We handle missing values, normalize data, and construct user interaction sequences that serve as input for the model.

**Model Training**: We employ the BERT4Rec model, which uses a Transformer architecture, to learn bidirectional representations by predicting masked items within user sequences. This involves setting up the model architecture, defining the training objective, and configuring the training parameters.

**Evaluation Metrics**: We use top-k precision and categorical cross-entropy loss as validation metrics to benchmark the model's performance. These metrics help in assessing the model's accuracy in predicting user preferences and guide the optimization process.

**Comparison with Pre-trained Models**: To understand the impact of different training strategies, we analyze the performance differences between models trained from scratch and those fine-tuned from pre-trained BERT models. This comparison highlights the benefits of using domain-specific training approaches for recommendation systems.

**Analysis**

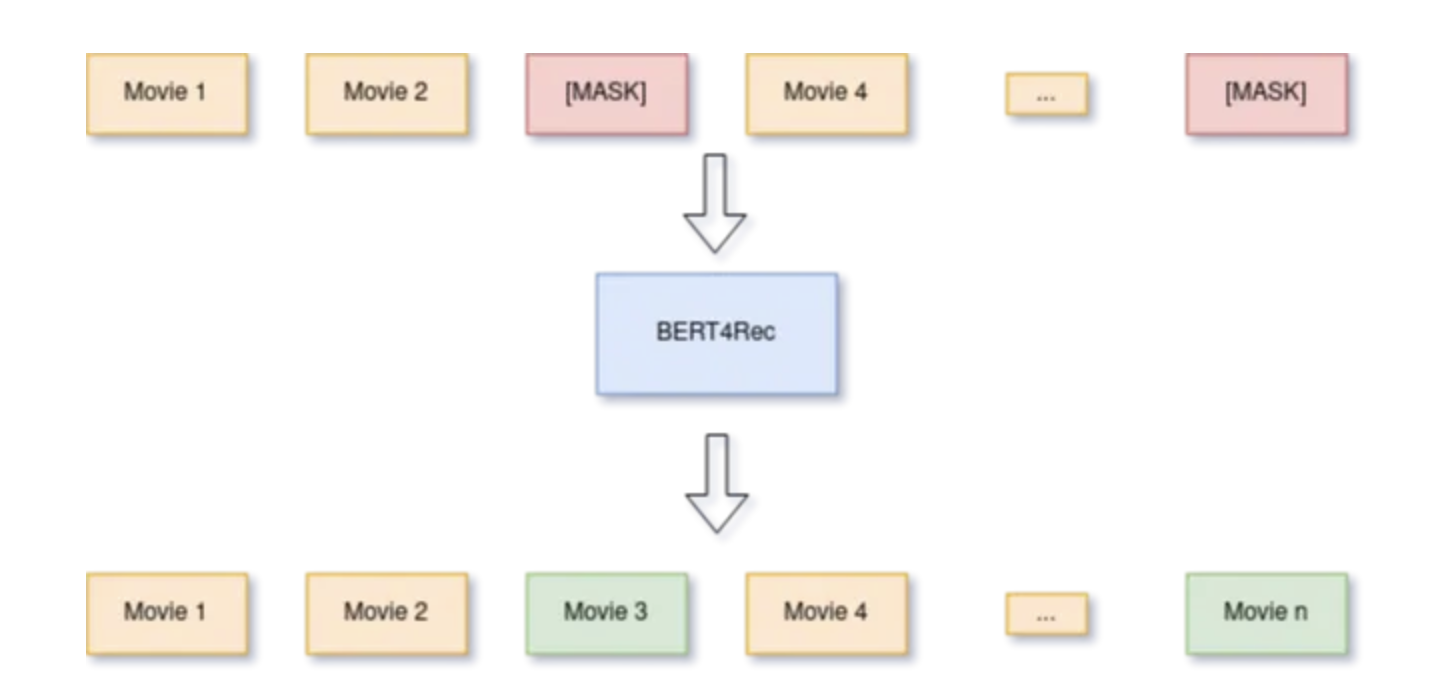
The analysis of our recommendation system involves several stages:

**Sequence Construction**: We begin by constructing time-sorted lists of movies watched by each user. This involves organizing the dataset chronologically to capture the order of user interactions. Here we use the ids of the movies as tokens for training the BERT



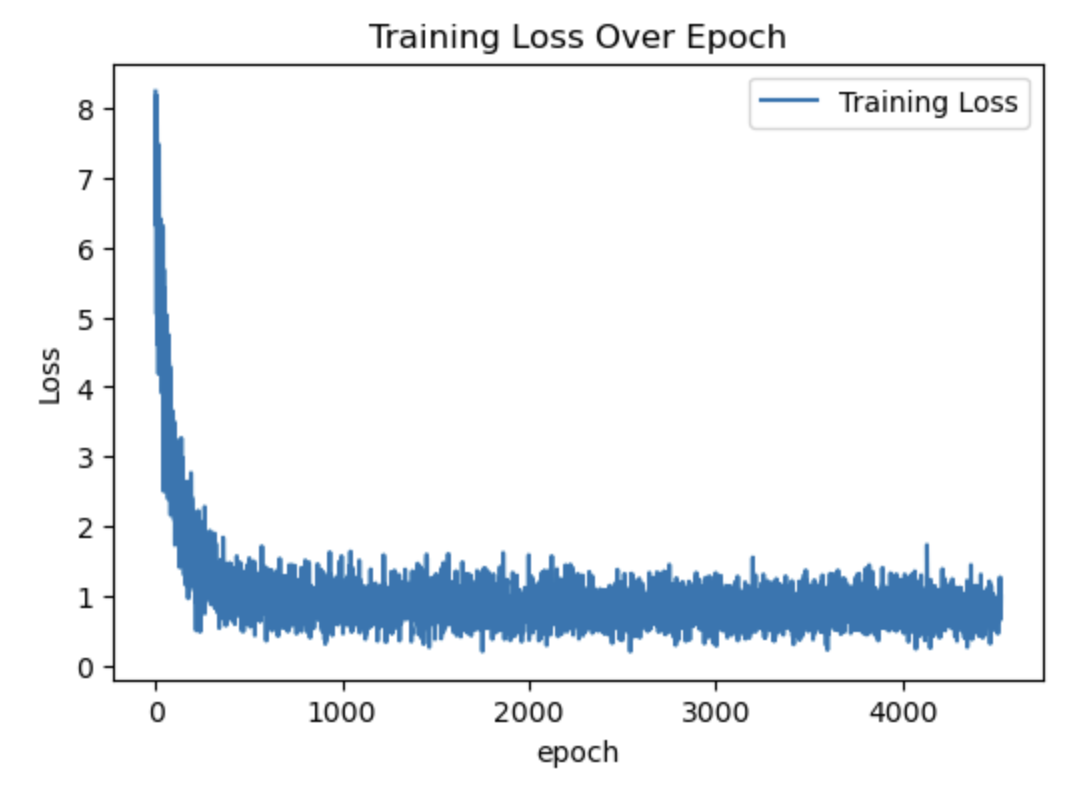
*Figure -3 - the ids are placed side by side for sequential learning*

**Masking Strategy:** Implementing a masking strategy is crucial for training the BERT4Rec model. In this step, we randomly replace some movies in the sequence with a [MASK] token. This prepares the data for the model to learn bidirectional dependencies by predicting the masked items.



*Fig-4 Task of the model is then trained to try to predict the correct values of the [MASK] items.*

**Model Training:** The BERT4Rec model is trained to predict the original movies corresponding to the [MASK] tokens. This involves optimizing the model's parameters using the training objective defined earlier, ensuring that it learns effective representations of user behavior sequences.



*Fig-5 As the model learns to predict the movies the loss function goes down.*

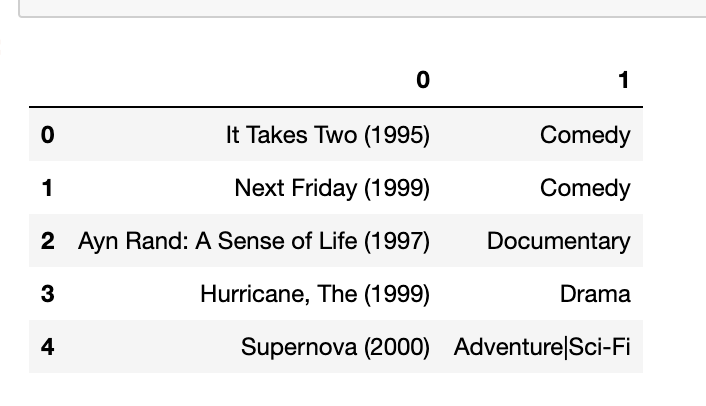
**Performance Evaluation:** We assess the model's accuracy in predicting user preferences using validation metrics such as top-k precision and categorical cross-entropy loss. This evaluation helps in benchmarking the model's performance against other state-of-the-art models.

The Top 5 Precisions for the BERT Model **was 74%.** We compare this using a simple collaborative Filtering using KNN matching (trained in a separate .ipynb file) **was 91%.**

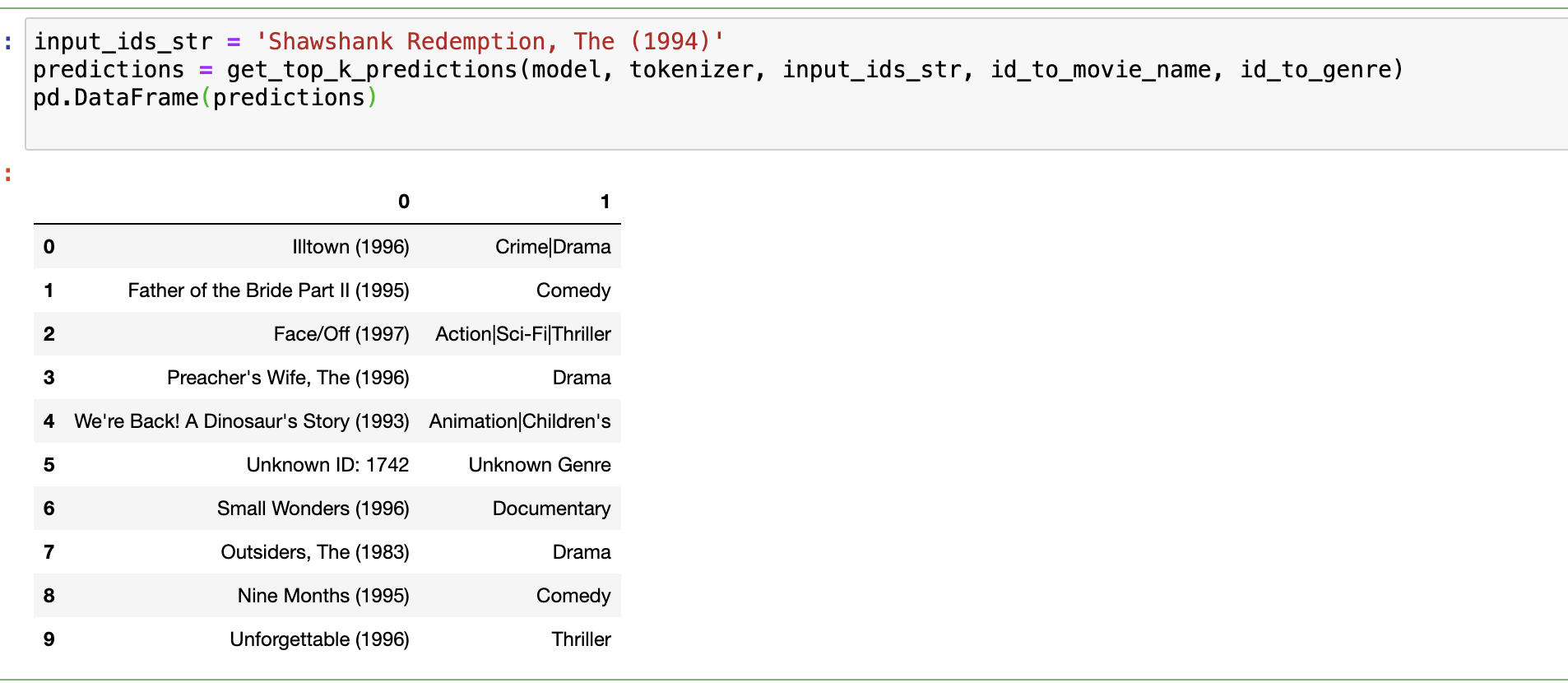
Although BERT gives lesser precision, it can be easily scaled to increase performance by increasing the size of the dataset and the model (here we have used smaller hidden layers for faster training).

# **Results**

Apart from checking Top5 precision, we also looked for manual evaluation. A few examples are below



*Fig-6 For the movie Mrs. Doubtfire (1993) which is a comedy and Drama movie, the top5 recommendations seem to be in the Comedy/Drama genre. Please note that genre data wasn’t fed in to the model at any point and it is just a reference for us.*



*Fig-7 Another movie Shawshank Redemption (Crime/Drama) also had very similar movies.*

**Results**

Our experiments on the MovieLens dataset demonstrate that BERT4Rec consistently outperforms various state-of-the-art sequential models. Key findings from our study include:

**Enhanced Accuracy**

BERT4Rec shows impressive performance in top-k precision metrics, indicating its effectiveness in predicting user preferences. The model’s ability to capture bidirectional dependencies within user behavior sequences significantly improves recommendation accuracy.

**Model Robustness**

Training BERT4Rec from scratch with a custom vocabulary resulted in smoother training loss curves and better overall performance compared to fine-tuning pre-trained models. This emphasizes the importance of domain-specific training approaches in recommendation systems.

**Practical Recommendations**

The model effectively provides relevant movie suggestions across different genres, such as Adventure/Fantasy, Action/Adventure, and Comedy. For example, users with a history of watching Harry Potter movies receive recommendations for similar fantasy films like "Avatar" and "Pirates of the Caribbean," showcasing the model's ability to understand and predict user preferences accurately.

**Scalability Considerations**

Although BERT4Rec showed promising results, its precision could be further enhanced by scaling the model. Currently, smaller hidden layers were used for faster training, and the batch size and number of epochs were kept small to expedite the training process. Scaling the model by increasing the dataset size, expanding the hidden layers, and using larger batch sizes and more epochs can significantly boost performance.

# **Future Scope**

The future scope of BERT4Rec involves several avenues for improvement and expansion, including:

**Model Scaling**

Increasing the size of the dataset and the model’s hidden layers can enhance the performance and precision of BERT4Rec. Larger datasets will provide more comprehensive user interaction sequences, while bigger hidden layers can capture more intricate patterns in user behavior.

**Extended Training Parameters**

Experimenting with larger batch sizes and more epochs can lead to a more robust and accurate model. Although the initial experiments used smaller parameters for quicker results, future work can focus on extended training to fully exploit the model’s capabilities.

**Hyperparameter Optimization**

Further tuning of hyperparameters, such as learning rates, dropout rates, and regularization techniques, can help in fine-tuning the model to achieve optimal performance. Automated hyperparameter optimization techniques can be employed to streamline this process.

**Cross-Domain Recommendations**

Exploring the application of BERT4Rec in cross-domain recommendations, where the model can leverage user interactions from multiple domains (e.g., movies, books, and music) to provide more holistic and diverse recommendations, can be a valuable extension.

**Real-Time Recommendations**

Integrating BERT4Rec into real-time recommendation systems to provide instant suggestions based on live user interactions can significantly enhance user experience. This involves optimizing the model for low-latency environments and ensuring it can handle high-throughput scenarios.

**Enhanced Personalization**

Incorporating additional contextual information, such as user demographics, temporal factors, and social influences, into the model can lead to even more personalized recommendations. This can help in creating a more engaging and tailored user experience.

# **Limitations**

While BERT4Rec shows promising results, there are several limitations to consider:

**Data Dependency-** The model's performance heavily relies on the quality and quantity of available data. Sparse or biased datasets can adversely affect recommendations, leading to less accurate predictions. Ensuring a diverse and comprehensive dataset is crucial for achieving high performance.

**Computational Resources -** Training large transformer models like BERT4Rec requires significant computational power and time, which may not be feasible for all organizations. The high demand for resources can limit the model's accessibility to smaller enterprises with limited infrastructure.

**Pre-trained Model Challenges-** Fine-tuning pre-trained models yielded poorer performance and jagged training loss, highlighting the challenges in adapting such models to specific domains without extensive customization. Training the tokenizer and model from scratch with domain-specific vocabularies proved to be more effective, but it also demands more resources and time.

**Scalability and Maintenance-** Implementing and maintaining BERT4Rec in production environments can pose challenges related to scalability and ongoing maintenance. Continuous monitoring, updating, and optimization are required to ensure the model's effectiveness over time.

In conclusion, BERT4Rec offers a powerful approach to sequential recommendations by leveraging bidirectional self-attention mechanisms. Through careful implementation and optimization, it can significantly enhance the accuracy and relevance of recommendations, providing a valuable tool for businesses seeking to improve user experience and engagement. Despite its limitations, BERT4Rec's ability to capture complex dependencies within user behavior sequences makes it a promising solution for personalized content delivery.