Example projects from HT2023

Marcel Bollmann

This document contains *selected* titles and abstracts from project reports submitted in HT2023. Only projects that received a passing grade are included. The projects are sorted alphabetically by their title and given a `badge` to indicate roughly which area(s) of Text Mining they correspond to.

Beyond the Game: Analyzing the Correlation Between NBA Players' Social Media Sentiments and On-Court Performance

`Text Classification` `Sentiment Analysis`

This study examines the potential relationship between NBA players' sentiments expressed on Twitter and their on-court performance statistics from 2017 to 2022. Utilizing sentiment analysis techniques on tweets from the top 15 NBA players, combined with an examination of their performance data, study aimed to explore the correlation between emotional expression on social media and professional basketball performance. Despite an intuitive hypothesis suggesting a potential reflection of players' sentiments in their game performance, the findings indicate a negligible correlation between tweet sentiments and performance metrics such as the Effective Field Goal Percentage (EFG). The Pearson correlation coefficients and p-values obtained from the analysis suggest that the observed relationship between these variables is statistically insignificant. This research contributes to the broader discourse on the impact of social media on professional athletes, highlighting the complexities of correlating psychological states with performance outcomes. Furthermore, it opens avenues for further investigation into the multifaceted effects and nuances of digital expression on physical performance in sports, suggesting that while direct impacts may not be observed, the role of social media remains a rich field for exploration.

Claim Verification Using Generated Data by ChatGPT and Wikipedia

Text Classification

Claim verification is the task of predicting a text's truthfulness. Verification models are often trained on manually created datasets which are expensive and time consuming to create. In this project, a verification model was trained on generated data. The data was generated by feeding scraped Wikipedia articles together with instructions to generate a true or false statement in a prompt to ChatGPT 3 turbo (via API). The articles were used as evidence, providing a source of truth for the model. The ChatGPT response was used as claims, either a true statement meaning it aligns with the evidence, or a false statement, meaning it contradicts the evidence. The evidence and claim were then embedded by using BERT-small. A neural network containing bidirectional LSTM layers was trained to distinguish between false and true claims. The model was able to classify the validation samples with a macro average F1-score of 0.80. After evaluating the model using custom inputs it was found that it is sensitive to the term 'not'. The reason for this could be due to an overrepresentation of the term 'not' in the false claims.

Comparing synthesized data techniques for class balance in cyberbullying prediction

Text Classification

This project uses machine learning models to analyze real world tweets in order to determine if the tweet contains cyberbullying. The project is mainly focused on unbalanced datasets, which means that the classes which constitutes the dataset are different in numbers. Unbalanced datasets is common when working with dataset generated from the real world, e.g. social media posts, in text analysis. Unbalanced datasets is usually a challenge for developers, since machine learning models tend to be biased for the class with the most samples. In order compensate for this, different machine learning models can be used to make the classes more even. This project introduces the reader to some of these models — SMOTE, kNN-SMOTE and undersampling. The result shows that kNN-SMOTE is a viable technique for data synthetization over SMOTE and undersampling.

Does Elon Musk's Twitter Account Affect Tesla Stock Prices?

Text Classification

This paper aims to prove or disprove a correlation between the tweets made by Elon Musk, CEO and founder of Tesla, and the change in price of Tesla stocks. This is done using sentiment analysis on tweets made by Musk during the period of 2017 through 2022 to determine if a positive tweet will make the prices increase and vice versa. Using the NLTK library for this technique and metrics such as mean square error and binary classification accuracy the results turned out worse than that of just random chance. Using the inverse of the binary classification metric, meaning that a positive tweet correlated to a decrease in price, a more promising result was reached. Using such a metric with an simplified investment algorithm resulted in a profit of 5.78 percent over a five year period. This might however be attributed to overfitting and not a true correlation.

Domain-specific QA using RAG

Information Retrieval Question Answering

In recent period, there has been significant advancement in the area of Question Answering (QA) systems, driven by the integration with powerful machine learning models and large-scale datasets. This paper explores the approach of Retrieval-Augmented Generative (RAG) models for the QA task by leveraging the power of both generative language models and information retrieval techniques. The paper begins by providing an overview of how a generative model performs on a QA task and its limitation, emphasizing the need for additional knowledge sources and improved methodologies to extract context from these sources to effectively answer complex queries. Our research delves into the intricacies of RAG and explores its impact on domain-specific QA tasks. Furthermore, we conduct comprehensive experiments in evaluating the RAG model using custom designed dataset and evaluation metrics that evaluates all components of the RAG model. The result demonstrates the effectiveness of RAG in achieving higher accuracy and context-aware responses.

Employing Text Mining for distinguishing American and Non-American movie categories

Text Classification

This project aims to use machine learning to sort movies into American and non-American categories, aiming to uncover cultural storytelling patterns in their plots. Through extensive exploratory data analysis and the application of various models, including logistic regression, support vector machine, and random forest, the analysis achieves notable success. By extensively analyzing the data and applying

various models like logistic regression, support vector machine, and random forest, the project succeeds notably. Specifically, the logistic regression model, considering word combinations of 1-3 grams, achieves an impressive F1-micro score of 0.864. This outperforms the basic models, indicating better precision and recall. The project demonstrates that machine learning can effectively distinguish cultural differences in movie narratives. The success of the logistic regression model, especially in analyzing word combinations, highlights its ability to understand and classify cultural patterns in diverse storytelling styles. This achievement holds promise for improving our grasp of cultural nuances in movies and shows the potential of machine learning in recognizing and categorizing these patterns.

Evaluating Automatic Code Summarization via Llama 2

Text Summarization

The study explores the premise of utilizing quantized Llama 2 models to achieve comparable results to ChatGPT, while significantly reducing computational requirements. The performance of the Llama 2 model is assessed using the ROUGE-L metric, and comparisons are made with results from previous studies. Findings indicate promising capabilities of the Llama 2 model in generating comprehensible code summaries, with potential for further improvements through additional evaluation metrics and parameter variations. The paper concludes with the authors insights into future prospects for enhancing automatic code summarization techniques.

Evaluating similarities in parties and the allegedly harshened debate

Text Classification Sentiment Analysis

The purpose of this project is to look at the sentiment of speeches made to the Swedish parliament using a BERT multi-label sentiment classifier. The project also attempts to use LDA to analyze motions proposed by different parties to determine which parties were most similar. All data was gathered from the Swedish parliaments open data. The project found a periodic increase of negative sentiment around elections. The study was not able to find any sufficient progress in determining similarity of parties, but instead analyzed the shortcomings and lessons learned.

Evaluation of various text classification methods in disease type detection based on symptoms

Text Classification

This project focuses on classifying text to identify disease types based on symptoms, employing various techniques such as eXtreme Gradient Boosting (XGB), Random Forest (RF), Multinomial Naive Bayes (MNB), Artificial Neural Network (ANN), and Convolutional Neural Network (CNN) for text classification. The Symptom2Disease dataset comprises 1200 data records with 'label' and 'text' components, is vectorized using a tf–idf vectorizer. Results show that the F1-scores range from 0.88 to 0.98, with the CNN model outperforming the others. Despite that CNN demonstrates superior performance, it has limitations, including higher computational cost, model building complexity, and intricate hyperparameter tuning. Future works are encouraged to utilize alternative models, e.g., Bidirectional Encoder Representations from Transformers, Recurrent neural network, etc.

Exploring Email Content with Topic Modeling and Embeddings

Topic Modelling

In the fast-paced world of business communication, managing and organizing a large number of emails is essential. Systems for classifying emails are essential for handling this deluge of correspondence. This study analyzes a collection of emails using two methods. Initially, we utilize BERT-based clustering to reveal hidden patterns. Secondly, we perform topic modeling using Latent Dirichlet Allocation (LDA) to uncover latent patterns in the emails. I hope to gain a better understanding of email communication in the hectic business world by integrating these approaches.

Exploring Global Discourse Evolution: Applying BERTopic, NMF, and LDA to the United Nations General Debate Corpus in the Pre/Post-Millennium Development Goals Era

Topic Modelling

We studied the United Nations (UN) General Debate Corpus, a collection of 7,314 speeches from 1970 to 2014. We wanted to understand how the focus and emphasis of speeches at the UN General Assembly changed after the adoption of the Millennium Development Goals (MDGs). We employed topic modeling techniques to identify key themes and topics in speeches. Our analysis revealed that BERTopic, a neural topic modeling algorithm, generated the most coherent topics. BERTopic's effectiveness in handling

complex datasets was evident. However, we also encountered challenges. BERTopic relies on pre-trained word embeddings, which may not effectively capture domain-specific information. Additionally, BERTopic can struggle with noisy data. Our dataset presented unique challenges, as it included scanned documents and complete text, deviating from the standard format of UN General Assembly speeches. The model's results indicated a degree of semantic similarity, but interpreting the results proved difficult.

Exploring Repetitiveness and Similarity of Song Lyrics Using Topic Modeling and Compression

Topic Modelling

Music nowadays is available to everyone, everywhere. But has appealing to a larger audience made today's hit-song lyrics more repetitive and similar than before? This study explores this question by using topic modeling and compression analysis to compare lyrics from every decade between 1960 and 2020. The lyrics are extracted from the Billboard Hot-100 charts. Results show that songs have become increasingly predictable and similar, although many of the common components have been there since the 60s. Songs can also on average be more compressed in recent decades than before, up to a maximum of 98.3%.

Extracting Topics from Swedish Parliament Speeches Using BERTopic

Topic Modelling

In the Swedish parliament, thousands of speeches are given every year. Representatives from all the parties in the parliament talk about important topics regarding our society. But what are they actually talking about? And how have the discussed topics changed over time? In this project, speeches held in the Swedish parliament on election years between 2010 and 2022 are analysed with the Topic Modelling technique BERTopic. Speeches were gathered from the Swedish parliament website, and after some pre-processing, fed to the BERTopic model. The results showed that 'family politics' regarding parents and children together with school politics were the most discussed topics during the entire time period, followed by animals' rights and women's rights. The model's performance was measured using both human judgement and coherence metrics using C_v and U_{mass}. The coherence metrics serve as a benchmark meant for comparison if further improvements and contributions are to be made to this project in the future. The manual inspection showed that the

model did not do a good job regarding placing similar topics in the same cluster; for example, electricity politics got two separate topics, and likewise with migration.

Injecting Domain Knowledge in Language Models for Financial Sentiment Analysis

Text Classification Sentiment Analysis

This project studies whether domain knowledge needs to be injected into sentiment analyzers for financial sentiment analysis. The performance of a RoBERTa base model trained on a general corpus and a RoBERTa base model pre-trained on a general corpus and fine-tuned on a financial data set are compared. The FiQA financial data set is used to evaluate the results. Although the domain-specific model outperforms the general-purpose one, the improvement is not that significant to justify the use of the domain-specific model in financial sentiment analysis.

Leveraging Large Language Model for Bias Detection in News Articles

Text Classification

In this study, a machine learning approach is developed to detect and measure bias in news articles. The methodology involves fine-tuning a language model specifically for the task of bias classification. To ensure robustness, the model is trained using a 5-fold cross-validation technique and oversampling methods. The model achieved a precision of 84% and 83% for SG1 and SG2 datasets respectively, ultimately aiming to minimize False Positive results in classifying biases.

The project categorizes bias into three types: distortion, content, and decision-making biases, and discusses their impact on media and democracy. The successful implementation of this project could lead to the development of tools that enable readers and journalists to better understand and check for biases, thereby contributing to a more informed and balanced public discourse. The paper provides a comprehensive discussion of the models used, the data collected, the evaluation methods employed, and the results obtained, offering an alternative to automated bias detection in journalism.

Leveraging LoRA to Compare Domain-Specific News Summary Generation: Domain Model vs. Larger General Model

Text Summarization

The burgeoning field of abstractive summarization employing large language models holds great promise, yet the associated challenges of time and resource-intensive in training for these models are substantial. To address the issue of the difficulty in training large models, various Parameter-Efficient Fine-Tuning (PEFT) technologies, including Low-rank adaptation (LoRA), have been introduced. This research encompasses two distinct tasks. Task A investigates the efficacy of fine-tuning the pre-trained model flan-t5-small using LoRA for enhancing its performance in generating news summaries. Task B explores the use of LoRA to ascertain whether larger general model outperform specialized-domain model in specific domains' task. This study proposes a research method inspired by the problem posed in a previous paper. The outcomes of Task A reveal significant improvements in news summary generation when employing LoRA for fine-tuning the pre-trained model. In Task B, it is observed that larger general model exhibits superior performance in specific domains compared to specialized-domain model. Evaluation metrics such as ROUGE and BERTScore are employed in both tasks, considering both text overlap and semantic similarity to provide a comprehensive evaluation of the models' performance.

Machine Learning Approaches to SMS Spam Detection

Text Classification

This project tackles the escalating issue of SMS spam, which poses a significant challenge to both mobile users and service providers. With the widespread adoption of mobile technology globally, the rise in SMS spam has become a major concern. By leveraging machine learning algorithms like the Support Vector Classifier, Random Forest Classifier, and XGBoost Classifier, alongside text mining techniques such as TF-IDF and Word2Vec, the study aims to effectively differentiate between legitimate messages and spam. Utilizing a dataset comprising 3068 messages collected from personal mobile inbox, the research assesses the performance of various models in SMS spam detection. Results highlight the Random Forest Classifier, Support Vector Classifier, and XG Boost Classifier as strong performers with TF-IDF representation, demonstrating high accuracy and ROC-AUC scores. Conversely, the baseline Random Classifier falls short, emphasizing the need for sophisticated models in text classification needs. The study underscores the importance of employing advanced techniques to enhance SMS security and improve user experience in mobile communication realms.

MBTI Personality Classification based on BERT

Text Classification

This project fine-tunes four BERT-based text classifiers, and uses user-generated content on the internet to predict their MBTI personality types. It compares whether large language models have advantages over simply duplicating data for text data augmentation, and also explores whether classification models trained on datasets with a relatively narrow range of topics have generalizability. The results show that using large language models to paraphrase the original data does not provide advantages for data augmentation. The generalizability of the classification models is also poor. However, the model seems to be able to capture similarities between personality types.

Predicting Genre Based on Lyrics: The Impact of Stop Words and Lemmatization

Text Classification

Many streaming platforms, such as Spotify, are dependent on providing their users with new relevant music. To effectively achieve this, different songs must be organized based on similarity or distinctions, and lyrics provide one way to do this. This project aims to explore genre classification based on song lyrics. It will be focusing on comparing the performance of a Naïve Bayes classifier and a logistic regression classifier. It will also explore how the removal and lemmatization may impact the results and be compared to when these processes are not included. The results show that the Naïve Bayes performed better than the logistic regression overall in accuracy, precision, recall and f1-score, which was consistent with the cross-validation scores. The removal of stop words and lemmatization resulted in a higher performance compared to when these processes were skipped.

Question-answering using KB-SBERT for The Swedish Transport Administration

Information Retrieval Question Answering

This project aims to explore the possibility of implementing a question-answering model at Trafikverket, The Swedish Transport Administration. The main goal of the model is for a user to be able to ask question about Trafikverket, its organization and operations using a chatbot deployed on Trafikverket's website. KB-SBERT and a fine-tuned version of the same model were implemented with a dataset of 1784 unlabeled question-answer pairs. The models works by taking the users question and comparing it to all questions in the dataset using cosine similarity. The answer

connected to the question with highest similarity is then returned to the user. The accuracy for the base and fine-tuned model were 53% and 38% respectively. Because of the unlabeled dataset, common evaluation metrics could not be used. Instead, ChatGPT were given input in the form of paragraphs from Trafikverket's website and then asked to produce questions for each paragraph. The questions were then used as input in the models and the yielded answers were then manually checked if correct. Even though the performance of the models were subpar, it shows that a question-answering model can be implemented, but that other methods like generative models should be explored.

Question answering with Retrieval-Augmented Generation

Information Retrieval Question Answering

With the rising popularity of large language models, there has been increasing interest in extending their knowledge beyond what they have been trained on. A solution to this called Retrieval-Augmented Generation (RAG) is evaluated in this paper and compared to a more traditional TF-IDF approach. This is done by implementing a retriever and a language model setup along with a dataset of syllabuses of courses given at Linköping University. This dataset was created by scraping the information from the university website called StudieInfo. The evaluation consists of 22 factual questions about individual courses and a further 22 questions asking which courses are the most similar. The results are then rated by a human evaluator on a scale from 1 to 5 based on how relevant the context was, how faithful the answer was to the context, and the quality of the answer. Interestingly, while the more complex transformer-based retriever performed better than TF-IDF on the comparative questions, with an answer quality of 4.091 and 3.091 respectively, the simpler TF-IDF performed better on the factual questions, with an answer quality of 3.810 and 2.619 respectively.

Reading Between the Lines: A BERT-Based Book Rating Predictor

Text Classification

The rating-inference problem is the novel problem of predicting the rating attached to a certain review. In this report, it is attempted to read between the lines of book reviews using a fine-tuned BERT model, in order to predict the rating attached to a book review as accurately as possible. The BERT model was fine-tuned into a rating classifier on a goodreads dataset consisting of roughly 1.37 M review-rating pairs. The BERT classifier fared well, outperforming human level and model baselines. Classifying the exact rating was rather difficult with an accuracy of 0.644, but predicting the

underlying sentiment of the review was much more of a success. This lead to the conclusion that the BERT classifier indeed can read between the lines to the extent of what is reasonable, but not beyond the subjectivity of the reviewer.

Sarcasm Detection with Transformer Model Encoders

Text Classification

This paper approaches the problem of detecting sarcasm in a text. A text can be sarcastic on its own or given a context, both scenarios are considered and it can be seen as a binary classification problem. Sarcasm is something that could be hard to understand even in speech and it has been showed that humans often relies more on the intonation in a voice than the context to understand sarcasm. To approach a problem of sarcasm detection in text, where the intonation in a voice is non existent, yields a challenge where the text itself and a possible context is the only thing that can be relied upon. The problem is approached by using the encoder part of transformer models. The data used consists of tweets, reddit comments and news headlines collected from multiple sarcasm-annotated datasets. During the project due to hardware limitations, the amount of data that could be used in the model was heavily reduced from about 1 million data points to about 50000. This of course effected the result and compared to the baseline which was trained on about 1.5 million reddit comments, the model presented in this paper was still performing similarly.

Sentiment Analysis of Steam Reviews

Text Classification Sentiment Analysis

Sentiment analysis, crucial for analyzing online trends and e-commerce decisions, forms an essential part of how we use technology to understand human language. In this study, we examined Steam game reviews to assess different sentiment analysis models. Using data from the Steam API, the project aimed to determine whether reviews were positive or negative. We compared a basic method, the Multinomial Naive Bayes model, with two advanced methods, BERT and RoBERTa. The Naive Bayes model showed reasonable results, its accuracy reached only 75%. In contrast, BERT and RoBERTa, particularly when used with balanced (undersampled) data, performed significantly better. RoBERTa achieved an impressive F1 score of 0.86, and BERT scored 0.82. These results highlight the advanced models' superior capability in accurately classifying sentiments in reviews, especially in datasets where positive and negative reviews are not evenly distributed. This study underscores the effectiveness of modern techniques in sentiment analysis. Furthermore, a manual error analysis

revealed that sarcasm and human error were significant challenges, especially for the advanced models.

Sentiment Analysis on Airline Tweets with XLNet and BERT

Text Classification Sentiment Analysis

Teaching AI the emotions expressed within a text based solely on words and contextual word relations has become a growing subject in recent times, even emerging as a field of its own, referred to as sentiment analysis or opinion mining. Transformer-based architectures have also become increasingly popular for this particular task, displaying state-of-the-art qualities. Two such models, BERT and XLNet, are compared in this project to see which performs better on the task of labeling the sentiment of posts on the platform X (previously Twitter) that are directing feedback toward US airlines. A BiLSTM model has also been created as a point of comparison. Ultimately, the classifiers' accuracy scores ended up close to one another, with accuracies of 82.04% and 81.83% on the BERT and XLNet model, respectively, concluding with the BERT model's performance being slightly better.

Sentiment analysis on long sequence game reviews

Text Classification Sentiment Analysis

Due to limitation in sequence length, traditional transformer models are not able process long sequence text. With the introduction of specialized transformers such as Longformer this becomes possible. By scraping new data from game reviews a completely new comparison could be made to test the performance of transformer models. After fine tuning the comparison was made between BERT, XLNet and Longformer on sentiment analysis of long sequence game reviews. The models were tasked with determining negative, neutral or positive sentiment based on the score of the review. Although XLNet is an state-of-the-art model the strengths of transformers with long sequence could be shown with Longformer outperforming all the other models based on Avr. weighted F1 score.

Sentiment Analysis on Steam Game Review using RoBERTa: Does Sarcasm Matter?

Text Classification Sentiment Analysis

Sentiment analysis on customer reviews is a widely used Natural Language Processing (NLP) application. However, some types of reviews, such as video game reviews, might contain sarcasm, which makes the traditional NLP model less than ideal. This paper aims to explore the state-of-the-art RoBERTa model's performance on the "Steam Review Dataset", which Steam is one of the biggest online video distributor. Also, four fine-tuned versions of the RoBERTa model using this specific dataset are tested. The performance of the fine-tuned model does improve over the base RoBERTa model and gives a more balanced result. On the other hand, the manual analysis of this paper suggests that although fine-tuned RoBERTa models achieve higher performance, they are not directly associated with the improvement to counter the effect of sarcasm.

Social Media Sentiment and Videogame Review Scores – A Cyberpunk 2077 Case Study

Text Classification Sentiment Analysis

This Paper tries to shed some light on the question if social media sentiment can reflect public opinions about a game and if it aligns with game review scores. A case study is performed on the game Cyberpunk 2077, conducting sentiment analysis on a data set of over 300k YouTube comments and comparing it to a second data set of almost 700k Steam reviews. The findings indicate that YouTube comments are able to capture shifts in public opinion on a coarse-grained level and the study is able to showcase a positive correlation between YouTube comment sentiment and Steam review scores. However, more fine-grained patterns in review scores over time are not clearly mirrored in the YouTube comment's sentiment scores. The study shows that YouTube comments can potentially be used to augment game review data sets, but broader research and improved filtering methods are needed to achieve that.

Spam vs. Ham: An analysis of Email Classification using Machine Learning Models and Neural Networks

Text Classification

This project focuses on classifying spam and ham emails, testing different machine learning models, such as Multinomial Naive Bayes or Random Forest Classifier, and also testing Neural Networks. There is an analysis of the data used in the project to solve problems like unbalanced data or irrelevant characters. Machine learning model evaluations reveal pretty good results, with Multinomial Naive Bayes being the most accurate one before and after applying Grid Search and Cross-Validation. A Neural Network model is introduced, taking it from previous work and trying

to improve it to get better results. Testing with randomly generated emails by AI demonstrates Multinomial Naive Bayes as the most effective classifier, providing the best accuracy compared to the other models tested. Finally, it has been found that a possible cause that prevents correct predictions is the presence of specific spam-type characteristics in the ham emails.

Teaching Large Language Models to Infer Information from Baseball Play Descriptions

Information Extraction

Large language models traditionally struggle with tasks that require complex reasoning or implicit knowledge, such as numeric reasoning. This paper attempts to address this by defining a structured task related to the pattern of certain tokens in the input prompt, then training a model for that task using a mixture of fine-tuning and prompt engineering. Specifically, the FLAN-T5 model was fine-tuned to answer basic questions about the outcome of a baseball game given only a description of the plays that occurred within the game. Using a training corpus of play descriptions for 5,221 Major League Baseball games, the model was trained in two stages. The first stage trained the model to calculate the number of runs a team scored, and the second trained the model to return the winning team and final score of each game. While the resulting model provided responses in the expected format, it mostly failed to return correct information about the games. However, prompt tests using more powerful models like ChatGPT were successful.

Textual Patterns in Fictional Literature

Text Classification Topic Modelling

This study explores the capabilities of basic text mining methods in analyzing the textual patterns of fictional literature, with a focus on the 'Harry Potter' series by J.K. Rowling. by employing Natural Language Processing (NLP) models, experiments will predict the specific book from a page of text. Through this quan- titative approach, a further understanding of how the readers perveive the story is produced. Through specific data preprocessing, the study addresses the challenges in varying formats of the books. Key aspects include the impact of named entities on model accuracy and the distinction between textual content and character dialogue styles. The research evaluates models' performance in classification, clustering, and topic modeling tasks, demonstrating how models gravitate to environmental clusters rather than plots. Initial findings confirm the hypothesis that named entities significantly influence the

models, with subsequent iterations focusing on removing these entities to enable focus on textual content. The study's scope and methodology offer a base for future research, suggesting the potential application of more advanced techniques like Large Language Models. This research contributes to a deeper understanding of the relation between NLP and literary analysis, highlighting text mining in the context of fictional narratives.

TV-show script generation

Text Generation

The enormous usage increase of advanced language models like GPT-3 and GPT-4 by OpenAI has opened up new possibilities in automated content creation, notably in script writing for television. In this project, we explore the surface of these models to see if it is feasible, particularly by using GPT-2, an older, open-source version of the more popular variants like GPT-3, and with Gemma, a scaled-down open-source version of Google's newest LLM Gemini. First inspired by the episode "Deep Learning", which was co-written by ChatGPT and Trey Parker, blending humans and AI creativity into mainstream media. The exploration in this project aims to generate new episode scripts based on the linguistic patterns and narrative styles learned from previous episodes. The process of fine-tuning GPT-2 and experimenting with the newer Gemma-2B model had partial success. Gemma-2B was extremely resource demanding and the GPT-2 achieved in emulating the character's dialogue patterns but faced difficulties maintaining narrative coherence and accurately portraying character relationships. These findings highlight the current limitations of AI in creative writing tasks, such as script generation. However, the project also underscores the promising role of AI in the creative process.

Who sang the song?

Text Classification

This project serves the purpose of classifying artists based on their lyrics. A total of five different models were trained and tested, four of which were trained and tested on the same dataset but with different preprocessing. This to be able to analyze how lemmatization affects the classification. The result shows that many artists write in canonical form and that stop words contain useful information in the case of lyrics analysis. Based on the result, it is possible to draw conclusions about what parts of the lyrics are important. The result showed that four of the classifiers had quite even

performance when looking at precision for each artist when compared to each other, while the BERT classifier differed from these quite a lot.