March 2024

Predicting Gender of Author Using Large Language Models (LLMs)

Satya Uday Sanku
University of South Florida

Predicting Gender of Author Using Large Language Models (LLMs)

by

Satya Uday Sanku

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science
Department of Computer Science and Engineering
College of Engineering
University of South Florida

Major Professor: Seungbae Kim, Ph.D.
John Licato, Ph.D.
Shaun Canavan, Ph.D.

Date of Approval:
March 11, 2024

Keywords: Prompting, Custom Feedback, Zero-Shot, Fine-tuning

Dedication

This is dedicated to my dad and mom, Mr. Kesava Swamy Sanku and Mrs. Sree Valli Sanku, who have always been there for me; my friends Mr. Anudeep Chinta, Mr. Theophilus Amaefuna, Mr. Raghu Vamsi Mummaneni, Ms. Thanuja Pavani Satti, and every friend who supported me; my sister Ms. Lakshmi Anjana Sanku and my brother Mr. Swamy Rakesh Adapa, who always believed in me; and my major professor Dr. Seungbae Kim, who guided me through my Master's. A big thank you to all of you for your love and support.

Acknowledgments

I want to express my deep gratitude to Dr. Seungbae Kim for offering me the opportunity to grow, learn, and challenge myself on the path to becoming an accomplished engineer. Your consistent support, guidance, and encouragement throughout my graduate journey bring immense happiness and inspiration to my heart.

I extend my thanks to Dr. John Licato and Dr. Shaun Canavan for their roles as valuable members of my thesis committee.

Heartfelt thanks to my parents, Kesava Swamy Sanku and Sree Valli Sanku, and my sister, Lakshmi Anjana Sanku. Your unwavering support and inspiration give me the strength to strive for success, and I will forever hold you in high esteem.

Lastly, my appreciation extends to those who showed kindness during this phase of my life. I remember and cherish your words, which have played a significant role in shaping me into a better person.

Table of Contents

List of Tables

List of Figures

Abstract

The advent of text data from social media, blogs, movie reviews, and other textual sources has opened new avenues for research, particularly in the domain of Author Profiling. Author Profiling helps in Capturing the Stylistic features and also useful for analyzing the required elements in the written text. This Study addresses one of the tasks in Author Profiling which is termed as gender detection or Classification of Gender from Text. The main goal of this research is to obtain valuable and relevant gender characteristics that will accurately classify the Author's gender of a review extracted from an Anime Review website. This Research uses the current State of Art Large Language Models to Automatically Capture the Gender Differences. The data is processed through the proposed Method which uses both Custom Prompting along with Fine-Tuning in order to tweak some of the weights associated with Large Language Models(LLMs). Once the LLM gets Fine-Tuned, the Model is tested with unseen Review datapoints, subsequently the Testing prompt is modified over the testing process through feedback mechanism proposed in the testing phase. Also, the Error Analysis is demonstrated through the Feedback obtained from the LLMs. Furthermore, the model surpasses existing baseline methods in accuracy, as evidenced by comparative analysis. This study contributes to the broader field of author profiling by presenting an effective model for gender detection and a thorough error analysis, highlighting potential areas for future enhancements and practical applications.

Chapter 1: Introduction

1.1 Motivation

Predicting an author's gender based on textual analysis is a significant task within the realm of Author profiling. Author profiling task is an interdisciplinary field that merges linguistics, computer science, and psychology to analyze textual content and predict the demographic and psychological attributes of authors. The essence of author profiling is to deduce characteristics like age, gender, ethnicity, native language, and personality traits from the stylistic and thematic features of their writing. One key application of author profiling is in forensic linguistics, where it aids in identifying the authors of anonymous or disputed texts by matching writing styles to individuals with known characteristics [13]. This capability is particularly valuable in criminal investigations and cybersecurity, where discerning the authorship of threatening messages or identifying sources of misinformation can be crucial [10]. In the digital marketing sphere, author profiling enables companies to understand the demographics and interests of content creators, facilitating targeted advertising and content personalization strategies [12], [28]. By analyzing blog posts, reviews, or social media updates, businesses can tailor their marketing efforts to specific audience segments, enhancing engagement and conversion rates [28]. Moreover, author profiling has gained traction in social media analytics, where it helps in categorizing users based on their writing styles, thereby improving content recommendation algorithms and fostering enhanced user experiences [11]. Platforms can use these insights to recommend more relevant content, ads, and connections to users, based on inferred characteristics. Considering this, according to research over the past years, data-specific preprocessing techniques are being performed in order to extract psycholinguistic features [8] such as:

- *Word n-grams* - Sequences of multiple words used as features to analyze writing style and compositionality. Unigrams, bigrams and trigrams are commonly extracted as features for author profiling models [25], [20], [17].

- *Part-of-speech tags* - Assigning lexical categories like nouns, verbs, adjectives to capture stylistic syntax patterns that can indicate gender [18], [17]. Features based on frequency of nouns, proper nouns, pronouns, etc. have shown predictive power [17].

- *Syntactic dependencies* - Analyzing sentence syntax by mapping relationships between words. The structure and complexity of sentences and phrases used can reveal gender differences [8], [23].

- *Semantic content* - Latent topics and themes derived from the text using topic modeling algorithms can enable profiling. Differences in men and women's topical interests become differentiating predictive signals [12].

Additionally, stylometric features related to readability, vocabulary richness, formality of language, emotiveness and complexity of sentences are also extracted using textual analysis techniques [8], [14]. Statistical approaches capture some of the significant differences between male and female writing styles. But still identifying specific gender differences from text is difficult and an open research problem [8].

Supervised learning algorithms like SVMs, Random Forests and Neural Networks [17] then utilize these multifaceted psycholinguistic signals from preprocessed text to predict gender with reasonable accuracy. However challenges remain in mitigating social biases, overcoming data differences and the tendency to reinforce stereotypes [24]. More research is vital.

1.2 Problem Statement

The primary objective is to accurately identify the gender of the author of a text that has been written by a human, while ensuring that the text has not been generated by (Artificial Intelligence)AI.

In this Thesis, A New Approach is introduced which utilizes Large Language Model's (LLMs) in order to Automatically capture the gender differences in the Human written text. The Experiments were conducted using Online Data which is extracted from Japanese Website called MyAnimeList [4]. The Data Extracted from this website [4] contains Anime Review Text and the respective gender association to that particular reviewer's text. The Main Goal is to reduce multiple number of preprocessing steps and to provide better interpretation for some of the misclassified data points. The comparison with the baseline models shows that our proposed method have achieved higher average F1-Score.

Chapter 2: Related Work

In Chapter 1 we discussed about the underlying motivations and defined the problem that this thesis aims to address. Building on that foundation, it is now essential to delve into existing research and review the literature concerning gender prediction from textual data.

2.1 Gender Prediction Using Classical Machine Learning Models

Mamgain et al. (2019, 2020) investigate author profiling with a focus on predicting gender and language variety based on document analysis, utilizing the PAN 2017 dataset. In their 2019 study [15], they employ machine learning and deep learning classification models such as Logistic Regression, Random Forest, Bag-of-Words (BoW), LSTM-CNN, and parameter tuning. Their results demonstrate that BoW performed best for gender prediction with accuracies of 0.8123 (training) and 0.7889 (testing), while LSTM-CNN yielded the highest accuracies for language variety prediction with 0.833 (testing). In their subsequent 2020 study [16], they further explore gender prediction using Natural Language Processing (NLP) techniques and machine learning classifiers including Logistic Regression, Multinomial Naïve Bayes, Random Forest, and Support Vector Classifier (SVC). The findings indicate that Logistic Regression outperforms other classifiers, achieving an accuracy of 64% on the English dataset. Both studies highlight the importance of author profiling in security, forensics, and marketing, and suggest that future work could focus on enhancing the accuracy of classification models and incorporating deep learning approaches.

Vashisth et al. (2020) analyze gender classification using Twitter text data, comparing traditional machine learning techniques such as Logistic Regression (LR), Support Vector Machine (SVM), Multilayer Perceptron (MLP), Naïve Bayes, Random Forest, and XG-Boost with Natural Language Processing techniques like Bag of Words and Word Embedding (W2Vec, GloVe). The study found that word embedding models, particularly W2Vec combined with LR, significantly outperformed traditional models, achieving an accuracy of 57.14% for gender classification, compared to the baseline TF-IDF with LR model, which achieved an accuracy of 53.65% [28].

Pizarro et al. (2020) presents models for profiling fake news spreaders on Twitter and gender profiling at PAN 2019 and 2020. The models use Support Vector Machine (SVM) classifiers trained with character and word n-grams. For gender profiling, the team achieved the best performance with an average accuracy of 0.8805. For fake news spreaders, they obtained one of the top results with an average accuracy of 0.7775 for English and 0.8200 for Spanish [21].

Cheng et al. (2011) investigate author gender identification from text, focusing on short, content-free text commonly found in Internet applications. They propose 545 psycho-linguistic and gender-preferential cues along with stylometric features for this identification problem. Using three machine learning algorithms (Support Vector Machine, Bayesian Logistic Regression, and AdaBoost Decision Tree), they conduct extensive experiments on large text corpora, including the Reuters Corpus Volume 1, newsgroup data, and Enron email data. Their results indicate an accuracy of up to 85% in identifying the gender, with function words, word-based features, and structural features being significant gender discriminators [9].

2.2 Gender Prediction Using Deep Learning Models

Khan Mohammadi et al. (2020) introduce PGST, a novel polyglot text style transfer approach for gender domain, which can be applied in multiple languages. They use a pre-trained word embedding for token replacement, a character-based token classifier for gender exchange, and the beam search algorithm for extracting the most fluent combination of suggestions. Their method aims to deceive gender identification models by transferring text style. They applied their approach to both English and Persian corpora, defeating their proposed gender identification model by 45.6% and 39.2%, respectively, and obtaining highly competitive evaluation results. The gender identification model achieved an accuracy of 90% in Persian and 80% in English. The CNN layers in their model are utilized to capture stylistic features, while the LSTM layers are used to capture long-term dependencies in text data [17].

2.3 Gender Prediction Using Transformer Based Models

Abdul-Mageed et al. (2019) explore age and gender detection in Arabic social media using deep bidirectional neural networks and variations of BERT. They demonstrate the utility of multi-task learning (MTL) for these tasks, particularly with task-specific attention, and show that a single-task BERT model outperforms MTL models. They report tweet-level accuracy of 51.43% for age and 65.30% for gender, both surpassing baseline models [5].

Studies have explored various machine learning and deep learning approaches for profiling tasks such as gender and age detection. These studies focused on finding the best combination of predictive features and selecting appropriate supervised machine learning methods or deep neural networks with hyperparameter tuning for maximum accuracy. However, they faced limitations such as explaining why specific models perform well, requiring extensive preprocessing time, and lacking interpretability. To address these issues, particularly preprocessing and feedback for misclassified data, the use of large language models (LLMs) has been proposed. The rest of the thesis is organized as follows: Chapter 3 gives the details about the Proposed Method, Ablation Study and Experimental results in Chapter 4 and Chapter 5 respectively, Limitations & Ethics in Chapter 6, Conclusion & Future Work in Chapter 7 and finally Contributions in Chapter 8.

Chapter 3: Methodology

The primary goal of this thesis is to achieve minimal preprocessing of data and to incorporate a feedback mechanism during the testing phase. To accomplish this, we employ custom prompting and fine-tuning techniques. The details and the entire process of the proposed method will be discussed in the subsequent sections of this chapter.

3.1 Dataset Description

The dataset used in this study consists of anime reviews extracted from a Japanese website called MyAnimeList [4]. This website features a variety of anime shows and movies, each accompanied by multiple reviews. The primary reason for selecting this website is that the reviews are linked to the author's profile, which includes information about their gender, a critical aspect for this research. Additionally, MyAnimeList is an open-source website, allowing for easy extraction of reviews using the Python library BeautifulSoup.

To ensure that most of the reviews are written by humans, we collected data associated with anime movies or shows released before 2019. After scraping the data from the website [4], we observed a higher number of male reviews compared to female reviews. To balance the dataset, we randomly sampled the male reviews to match the number of female reviews. Initially, we extracted 60,000 reviews, comprising 40,000 from males and 20,000 from females. After sampling, the dataset comprised 40,000 reviews, with an equal distribution of 50% male and 50% female reviews. The dataset is represented in the form of Equation 3.1.

$$D = \{(d_i, y_i) \mid i = 1, \dots, n-1\} \tag{3.1}$$

In Equation 3.1, D represents the entire dataset, d_i represents the anime review text, and y_i represents the author's gender for the corresponding review text d_i, where y_i can be either Male or Female. The index i represents the position of each pair in the dataset, ranging from 1 to $n-1$, indicating the location of each data point within the dataset.

Table 3.1: Comparative Word Frequency Analysis by Gender in Reviews

	Male	**Female**
Total Reviews	23251	23251
Min Number of Words Used	23	15
Max Number of Words Used	9294	9537
Average Word Frequency	611	444
Median Word Frequency (W_f)	444	391
25th **percentile Reviews** W_f	235	185
75th **percentile Reviews** W_f	797	579

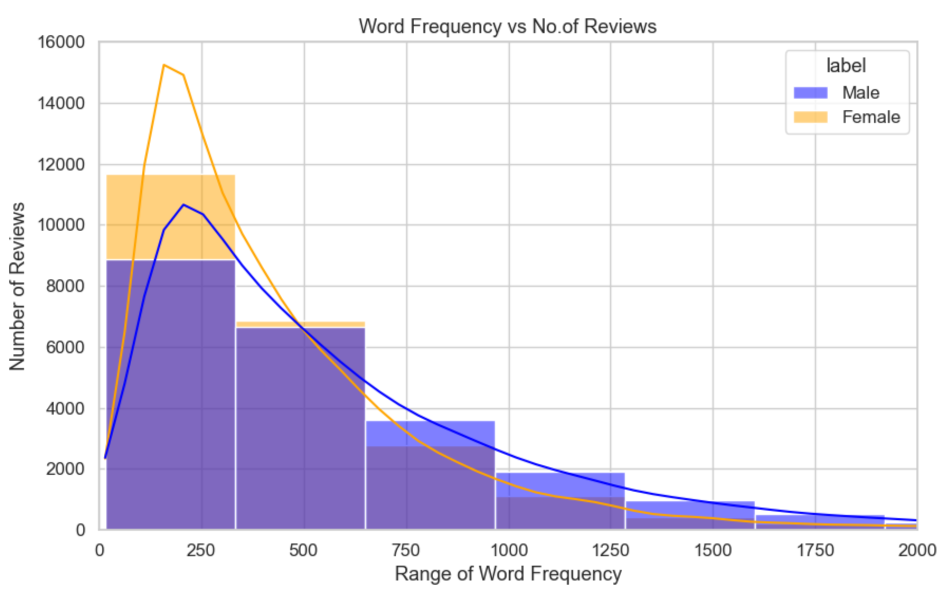

Figure 3.1: No. of Reviews vs No. of Words

The comparative word frequency analysis in Table 3.1 and Graph 3.1 shows that Male reviewers tend to write longer reviews, with a minimum of 23 words compared to female's 15, and a maximum of 9,294 words versus female's 9,537. The average word frequency for males is 611, higher than female's 444, indicating that men generally use more words in their reviews according to this particular dataset.

The median word frequency, a robust measure of central tendency, is 444 for males and 391 for females, further supporting the observation of more verbose male reviewers. Additionally, the distribution of word frequencies shows that the bottom 25% of male reviewers use at least 235 words, while the top 75% use at least 797 words, both higher than their female counterparts 185 and 579 words, respectively.

Overall, the analysis suggests that male reviewers consistently use more words in their reviews across various measures, including minimum, maximum, average, and median word frequencies, as well as the 25th and 75th percentiles. This difference in linguistic behavior may be influenced by factors such as communication styles, preferences, and social norms.

We will see how this data is used in this experimentation for further gender analysis and prediction in further sections.

3.2 Prompt Engineering

Prompt engineering is the process of designing and optimizing prompts to effectively guide large language models (LLMs) like GPT-3 in generating desired outputs for specific tasks. This involves crafting the input text or instructions in a way that maximizes the model's performance and accuracy [22]. In prompt engineering, the goal is to create prompts that are clear, concise, and aligned with the task at hand. This can involve experimenting with different formulations, structures, and levels of detail to find the most effective way to communicate with the model. The process often requires a deep understanding of the model's capabilities and limitations, as well as the nuances of the task being addressed [1]. Effective prompt engineering can significantly enhance the performance of LLMs, making them more useful and reliable for a wide range of applications, from natural language processing and content generation to complex reasoning and decision-making tasks [2].

Within the realm of prompt engineering, there are two notable types: Prompting and Chain of Thought Prompting

3.2.1 Prompting

Prompting involves providing a model with a specific input or set of instructions to guide it in generating a desired output. This method is straightforward, where the prompt directly asks the model to perform a task or answer a question. For example, a simple prompt might be, "Classify the sentiment of the following text as positive, negative, or neutral."

3.2.2 Chain of Thought Prompting

Chain of Thought Prompting, on the other hand, is a more sophisticated technique that guides the model through a series of logical steps or intermediate thoughts to arrive at a final answer. This approach is especially useful for complex tasks that require reasoning or multiple steps to solve. The prompt not only asks the model to perform a task but also to explain the thought process leading to the final answer, as demonstrated by Wei et al [29].

3.2.3 Example Prompt

The Chain of Thought Prompting approach is utilized to predict the gender accurately. Instead of directly asking the model to classify the gender, the prompt 3.2 guides the model through a thought process. It starts with analyzing the text, considering specific linguistic features, and identifying patterns related to gender. This structured approach helps the model reason through the task and provide a more informed and accurate prediction of the author's gender.

Prompt:
You are a linguistic analyst specializing in gender identification through written text.
Instruction 1
Instruction 2
Instruction 3
⋮
Instruction 5: Indicate the predicted gender with ‘Gender:’, which should be either

‘Male’ or ‘Female’. Remember, your response must be limited to these two options.

Figure 3.2: Prompt

3.3 Zero-Shot vs Fine-tuning

In natural language processing (NLP), zero-shot learning and fine-tuning are two distinct approaches for training models to perform tasks on unseen data. Zero-shot learning enables a model to make predictions on tasks or data it has not been explicitly trained for, relying on its general understanding of language and task descriptions. This approach is advantageous when labeled data is scarce or when the model needs to generalize across a wide range of tasks without retraining. Fine-tuning, on the other hand, involves training a pre-trained model on a specific task with a smaller dataset to adapt its knowledge to the task at hand. This method is effective in improving model performance on the target task by leveraging the rich representations learned during pre-training. While zero-shot learning offers greater flexibility and scalability, fine-tuning often results in higher accuracy for specific tasks, making it a preferred choice when task-specific data is available [7].

3.4 Approach

In this method, the Open AI GPT 3.5 will be used to Fine-tune the Anime Review text data by using the custom prompting

3.4.1 Data Preparation

For our study, we started with a dataset of 40,000 reviews. We split this dataset into two parts in which 30,000 reviews for training and 10,000 reviews for testing. Due to the limitations of GPT-3.5, which can handle only up to 16,385 tokens, we had to be careful with how we used our training data. We selected 2,000 reviews from the training set to fine-tune our model. Out of these, 1,700 reviews were used for the actual training process, while the remaining 200 reviews were used for validation purposes. This approach allowed us to efficiently utilize our data within the constraints of the model.

3.4.2 Fine-tuning

Upon collecting an adequate number of data points, specifically 2000 reviews in this case, the fine-tuning phase is initiated. This phase involves sending the reviews, along with a designated training prompt (Figure 3.3), to the OpenAI GPT-3.5 Turbo model [19]. The goal of fine-tuning is to adapt the model to better align with the specific requirements of the task at hand, thereby enhancing its performance.

Prompt:
You are a linguistic analyst specializing in gender identification through written text.
Instruction 1: Your task is to predict the gender of the author of an anime review, indicated by 'Review:'
Instruction 2: This prediction should be based on linguistic indicators such as word choice, sentence structure, themes, comments on musical elements, and emotional expression.
Instruction 3: Given the complexity of this task due to a variety of writing styles and the absence of definitive gender-specific linguistic patterns, your judgment should rely on a probabilistic analysis of gender-related tendencies in writing styles, especially in the context of anime reviews.
Instruction 4: Indicate the predicted gender with 'Gender:', which should be either 'Male' or 'Female'.
Instruction 5: Remember, your response must be limited to these two options.

Figure 3.3: Training Prompt

In this research, the Anime Review Text data, along with the gender labels, are used for fine-tuning to automatically capture gender-specific patterns and update the model's understanding accordingly.

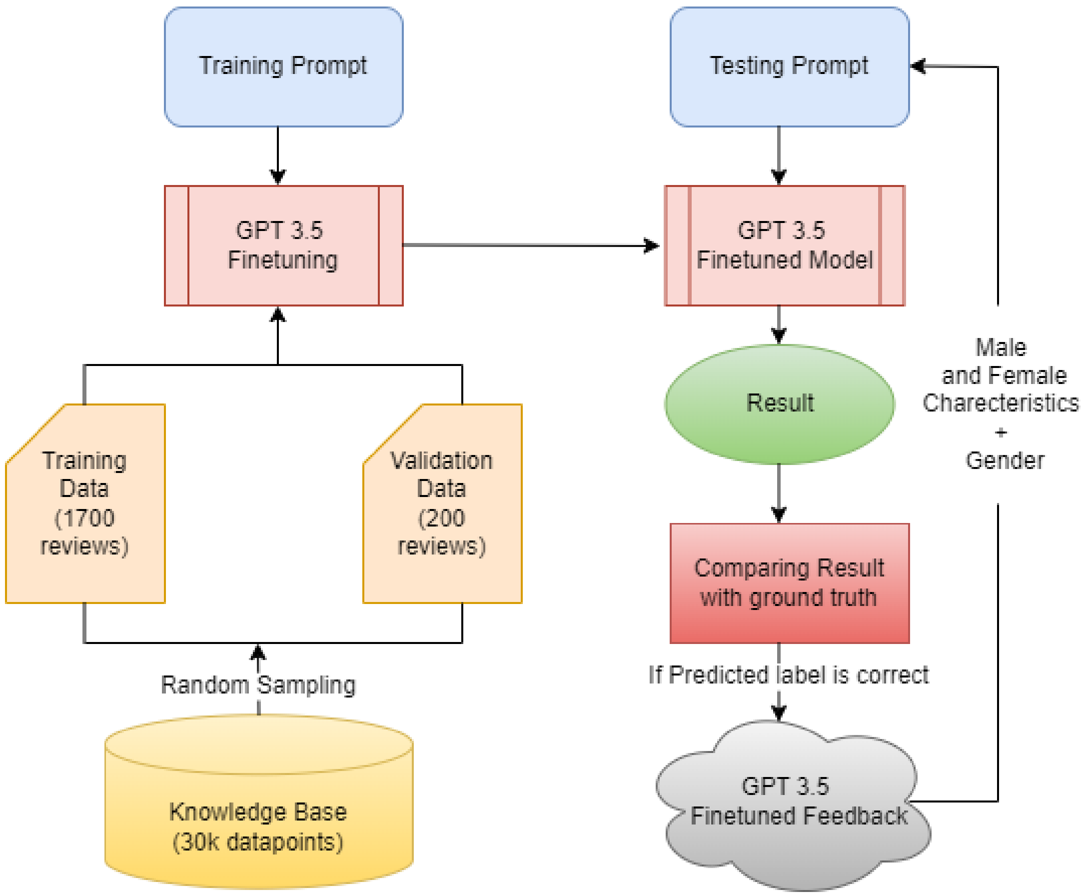

Figure 3.4: Our Method

The fine-tuning process for GPT-3.5 Turbo involves several steps (Figure 3.4). Initially, the data must be prepared in a specific format, typically as a JSON file, containing messages with roles such as "system", "user", and "assistant", along with their content. Next, the prepared data files are uploaded using the OpenAI API for fine-tuning purposes. A fine-tuning job is then initiated through the OpenAI API, specifying the training file and the model ("gpt-3.5-turbo"). Once the fine-tuning process is complete, the custom model can be employed to generate responses by making API calls that specify the fine-tuned model.

Fine-tuning offers several advantages, such as improved steerability, consistent output formatting, and customized tone, all of which contribute to better performance for the specific use case. After the completion of the fine-tuning process, a notification is sent via email, indicating that the model is ready for deployment. The fine-tuned model can then be integrated into applications or services, leveraging its improved capabilities to capture gender-specific patterns accurately and update its responses accordingly [6], [30], [3].

3.4.3 Testing Phase

Once the fine-tuning process is complete, the model enters the testing phase. During this phase, the model is presented with a testing prompt 3.5 that includes a review from an unknown author. The task of the fine-tuned model is to predict the gender of the author by analyzing the review text, leveraging the knowledge it has acquired during the fine-tuning process. In this research, the testing prompt shown in Figure 3.5 is designed to evaluate the model's ability to identify gender-specific patterns in the Anime Review Text data. The prompt is crafted to mimic real-world scenarios where the model might be deployed, ensuring that the model's predictions are both accurate and relevant to the task. The testing phase plays a crucial role in assessing the model's performance and its generalization capabilities to new, unseen data. It provides an opportunity to evaluate the model's strengths and weaknesses and to identify areas where further fine-tuning or adjustments may be necessary. A careful analysis of the model's predictions during the testing phase helps in understanding the model's proficiency in capturing gender-specific language patterns and its ability to apply this understanding to make accurate predictions [27], [6].

3.4.3.1 Incorporating Feedback

In the testing phase, the fine-tuned model is prompted with a series of review instances from the test set. For the initial review, no feedback from previous predictions is incorporated into the prompt as shown in Figure 3.5.

Testing Prompt:
User: Given the Review: "+row['review']+" predict the gender by using the finetuned knowledge, indicated by '**Gender:**'

Assistant: Gender: ""

Note: row['review'] represents the review text from the Testing Dataset

Figure 3.5: Testing Prompt without Feedback

Testing Prompt:
User: Given the Review: "+row['review']+" predict the gender by using the finetuned knowledge, indicated by '**Gender:**'

While Predicting the gender Consider the below Feedback of both Male and Female but you are not limited to these Characteristics

Male: "They often mention the animation quality and soundtrack as important factors"
Female: "focus more on the emotional aspects of the anime, such as overall message of the story"

Assistant: Gender: ""

Note: row['review'] represents the review text from the Testing Dataset

Figure 3.6: Testing Prompt with Feedback

However, starting from the second review, the model utilizes feedback from the previous review's prediction as shown in Figure 3.6, to enhance its analysis. If the previous gender prediction is accurate, the Finetuned model is given a new task to identify the characteristics based on which it classified the review as being written by a particular gender. This is achieved by embedding the previous review and its predicted gender into a prompt, which instructs the fine-tuned model to generate text that explains the characteristics that led to its classification. These generated characteristics, whether male or female, are then used to modify the testing prompt for subsequent reviews. This modification aims to leverage both the fine-tuned knowledge and the characteristics identified from the model's feedback, thereby improving the accuracy of the author's gender classification. By incorporating feedback from previous predictions, the testing phase becomes more dynamic and adaptive, allowing for continuous refinement of the model's performance over the testing data samples based on its successes and failures.

Chapter 4: Ablation Study

An ablation study is a research technique commonly used in the fields such as machine learning, Natural Language Processing, neuroscience, and biology [26]. It involves systematically removing or altering components of a system, process, or model to understand their individual contributions to the overall performance or function. In the context of machine learning, for example, an ablation study might involve removing certain features, layers, or parameters from a neural network to evaluate their impact on the model's accuracy or efficiency. This helps researchers and practitioners identify the most critical elements of a model and can guide further optimization or simplification efforts.

Table 4.1: Evaluating GPT-3.5 Performance Using Different Prompting Techniques

Finetuned	**Feedback**	**Static/Dynamic**	**Precision**	**Recall**	**F1-Score**
×	×	×	71.0	70.0	69.5
×	✓	Static	70.5	70.5	70.5
×	✓	Dynamic	71.0	70.5	70.5
×	✓	Dynamic$_f$	73.5	73.5	73.5
✓	×	×	78.0	77.5	77.5
✓	✓	Static	81.5	80.5	81.0
✓	✓	**Dynamic**	**82.0**	**81.5**	**82.0**

In this research, we conducted an ablation study to thoroughly investigate the effectiveness of our proposed method. This involved making systematic alterations to our model's prompting techniques and observing the impact of these changes on performance. We explored both zero-shot prompting, where the model makes predictions without prior specific training, and fine-tuning techniques (see Section 3.3), where the model is further trained on related tasks to enhance its accuracy. Through this methodical approach, we aimed to elucidate the strengths and limitations of our model under various conditions and refine our understanding of its behavior.

In the Table 4.1 the first row represents the model's performance without any finetuning, feedback, or prompting strategy which is zero-shot learning, where the model is evaluated on its ability to understand and respond to tasks it hasn't been explicitly trained on. Rows 2 and 3 introduce feedback into the prompting strategy, with Row 2 (Static Feedback) incorporating a fixed feedback mechanism, and Row 3 (Dynamic Feedback) allowing the feedback to adjust based on the input. Row 4 (Dynamic Feedback from Finetuned Model) explores the use of dynamic feedback sourced from a finetuned model, although the model performing the task remains unfinetuned.

Rows 5 through 7 focus on models that have undergone finetuning, which involves additional training on a specific dataset or task to improve performance. Row 5 (Finetuned without Feedback) presents the performance of a finetuned model without any feedback mechanism. Row 6 (Finetuned with Static Feedback) adds a static feedback component to the finetuned model. Additionally, from the Row 6 result we can see that the Randomly sampled feedback from the list of fine-tuned feedback's helps the model to distinctly classify most of the Male and Female reviews which shows that having a better static prompt along with task specific finetuning helps to improve the models performance.

Finally, Row 7 represents our Proposed Approach, where the finetuned model utilizes dynamic feedback in its prompting strategy. According to the results, this approach yields the best performance metrics, indicating that the combination of finetuning and dynamic feedback is beneficial.

The improvement in precision, recall, and F1-score in the last row highlights the effectiveness of our proposed methodology. This table demonstrates the incremental impact of different training and feedback strategies on the model's performance, with the proposed approach showing significant advancements over the baseline and other variations.

Chapter 5: Experimental Results

5.1 Experimental Settings

In this study, we employ various baseline methods to evaluate the performance of our proposed model. The baseline methods include Zero-Shot Prompting, a technique that leverages a pre-trained language model to make predictions without any additional fine-tuning. We also explore traditional machine learning models such as Logistic Regression, Naïve Bayes, and Decision Trees, which are trained on embeddings generated by the Roberta Base model. Additionally, a 4-layered Deep Neural Network (DNN) is utilized, with the network being trained using the same Roberta Base pre-trained embeddings.

The Roberta Base model, an optimized variant of the BERT model, is employed for generating textual embeddings. These embeddings serve as the input features for both the machine learning models and the DNN. The DNN architecture comprises four fully connected layers, with appropriate activation functions and dropout layers to prevent overfitting.

For training the models, we use the Cross-Entropy loss function, which is well-suited for multi-class classification tasks. The Adam optimizer is chosen for its efficiency in updating model parameters, and we employ the ReduceLROnPlateau scheduler to adjust the learning rate based on the performance on the validation set. The learning rate is reduced by a factor of 0.5 if there is no improvement in validation loss for 5 consecutive epochs. The models are trained with a batch size of 32, and we implement an early stopping criterion, where training is halted if there is no improvement in validation loss for 15 epochs, to avoid overfitting and reduce computational time.

To assess the performance of the models, we use the Average F1-Score as the evaluation metric. The Average F1-Score is a balanced measure that considers both precision and recall, making it suitable for evaluating models on imbalanced datasets.

By adopting these experimental settings, we aim to provide a comprehensive evaluation of our proposed model and baseline methods, ensuring a fair comparison and robust analysis of their performance.

5.2 Comparision Between Baseline Models and Our Method

Our method achieves an F1 Score of 82%, which is a significant improvement over the baseline models. Among the baseline models, the combination of Roberta Base and Decision Tree exhibits the highest performance, with an F1 Score of 71%. This indicates that our method outperforms the best baseline model by a notable margin of 11%.

Table 5.1: Results Comparison Table

Model Name	**Accuracy**	**Precision**	**Recall**	**F1-Score**
RoBERTa Base + Naïve Bayes	60	60.5	60.5	60.5
RoBERTa Base + Logistic Regression	70	69.5	69.5	70
Zero Shot with GPT-3.5	70	71	70	70
RoBERTa Base + Decision Tree	71	71.5	71.5	71
RoBERTa Base + Deep Neural Network	73	73	73	73
Our Method	**82**	**81.5**	**82**	**82**

When comparing our method to Zero Shot, a technique used in previous research, our approach shows a marked improvement. Specifically, our method retains approximately 10% more male data than Zero Shot. This is reflected in the F1 Scores for gender-specific data:

our method achieves an F1 Score of 81% for male data, which is 15% higher than the 66% achieved by Zero Shot. Similarly, for female data, our method attains an F1 Score of 83%, outperforming Zero Shot's score of 73% by 10%.

Overall, our method demonstrates a high average F1 Score of 82%, indicating its effectiveness in the task at hand. The improvement in performance is consistent across different subsets of the data, including both male and female data. These results suggest that our method not only achieves overall high performance but also maintains a balanced performance across different demographics. This is an important consideration in developing fair and equitable models. The substantial improvement over both the baseline models and Zero Shot highlights the potential of our method in advancing the state of the art in this domain.

5.3 Error Analysis Using Finetuned GPT 3.5

We conducted an error analysis to understand why some data points were incorrectly classified by our custom fine-tuned GPT 3.5 model. The initial prompt we used was to identify the distinct characteristics of males and females based on the knowledge fine-tuned into the model. The results of this analysis are displayed in Figure 5.1, which we've labeled as "Analysis of Distinct Characteristics."

Additionally, we examined the misclassifications of male and female data points. The analysis of errors in female data predictions is shown in Figure 5.5, while the analysis of errors in male data predictions is presented in Figure 5.4. In both of these analyses, we observe that male reviewers tend to use a majority of characteristics typically associated with females in their writing, and female reviewers tend to use a majority of characteristics typically associated with males. This crossover in writing styles leads to misclassification in our proposed Method.

The fine-tuned model generated a list of common words that are frequently used by males and females, which is presented in Figure 5.1. To verify that these words are actually used in the data that was fine-tuned for this gender prediction task, we plotted the word frequency

Prompt:
User: Based on the finetuned anime data, Identify the distinct characteristics in reviews of male and female?

Feedback from Finetuned Model:

Male Reviewers tend to focus more on the action, fight scenes, and overall plot of the anime. They often use words like "**action**", "**fight**", "**power**", "**battle**", "**cool**", "**epic**", "**intense**", and "**badass**" to describe their favorite moments. They often mention the animation quality and soundtrack as important factors in their reviews.

Female Reviewers tend to focus more on the emotional aspects of the anime, such as character development, relationships, and the overall message or theme of the story. They often use words like "**love**", "**cute**", "**romance**", "**heartwarming**", "**emotional**", and "**beautiful**" to describe their favorite moments.

Overall, male reviewers tend to emphasize the action and excitement of the anime, while female reviewers tend to emphasize the emotional and relational aspects.

Figure 5.1: Analysis on Different Gender Characteristics

in graphs, as shown in Figures 5.2 and 5.3. These graphs represent the percentage of reviews that contain a particular word, indicating how frequently these words are used across the entire dataset.

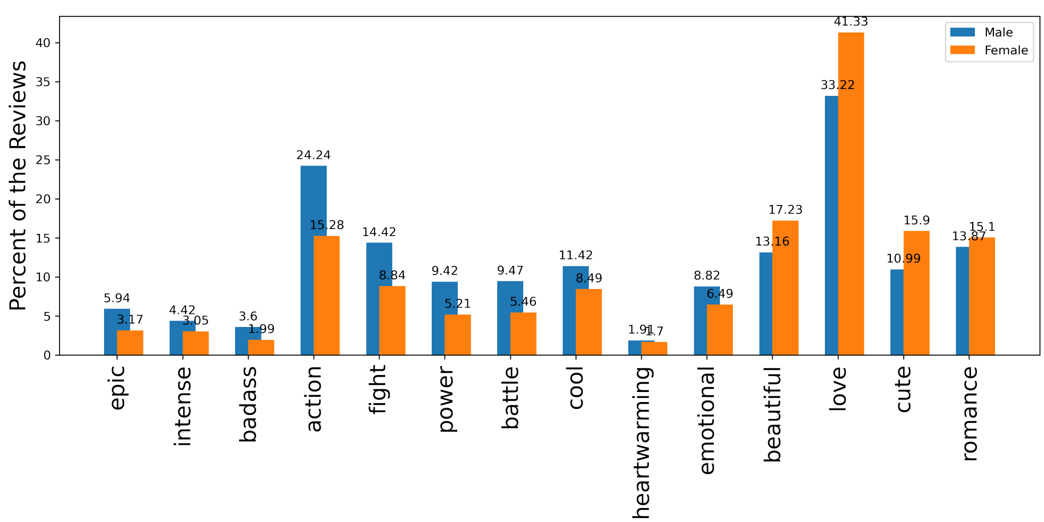

Figure 5.2: Word Usage Analysis in Fine-tuned Data

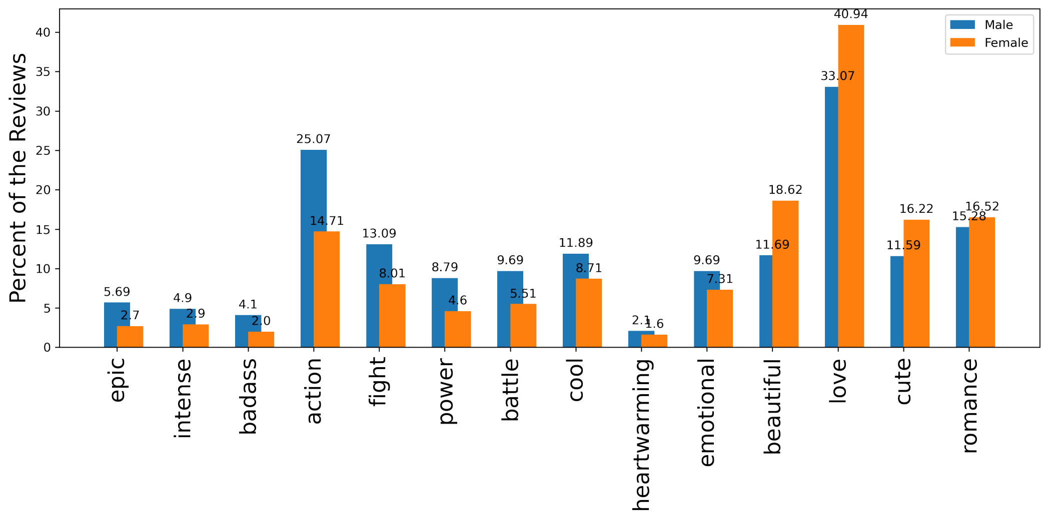

Figure 5.3: Word Usage Analysis in Fine-tuned Data Along with Test Data

Upon examining these graphs, it is observed that words like "heartwarming" and "emotional" are used more often by male reviewers than female reviewers. This indicates that the model may be exhibiting some bias in this case. However, the remaining words are correctly predicted with respect to the reviewer's gender, which indicates the model's capability to generalize correctly in this context.

Prompt:
User: Here are all the reviews identified with Review 1 to Review 18 +all_reviews+ List the Characteristics that are found common in all 18 reviews in detail (100-118 "Wrong predicted as Female")

Feedback from Finetuned Model:

- Personal opinions and experiences with the anime - **Female**
- Detailed analysis of story, art, sound, characters, and enjoyment – **Female/ Male**
- Emotional reactions to the anime, including enjoyment, excitement, and disappointment - **Female**

Figure 5.4: Error Analysis (Male Example)

Prompt:
User: Here are all the reviews identified with Review 1 to Review 18 +all_reviews+ List the Characteristics that are found common in all 18 reviews in detail (100-118 "Wrong predicted as Male")

Feedback from Finetuned Model:

- The reviews are all focused on anime, manga, or Japanese culture.
- The reviews express personal opinions and subjective viewpoints. -**Female** subjective viewpoints - **Male**
- The reviews contain a mix of positive and negative feedback. - **Male/ Female**
- The reviews may contain spoilers or specific plot details related to the anime or manga being discussed. - **Male**
- The reviews may contain humor, sarcasm, or informal language. - **Male**

Figure 5.5: Error Analysis (Female Example)

Chapter 6: Limitations and Ethics

6.1 Limitations

In this research, the proposed method uses anime data for fine-tuning purposes, which limits the model's performance to this particular review dataset. To generalize the model to other datasets, there is a need for a wider variety of datasets to fine-tune the model. However, due to the maximum token limitation for fine-tuning GPT 3.5, a larger number of data points cannot be incorporated to tweak the model during the fine-tuning phase. This constraint poses challenges for achieving broader generalization of the model.

If reviewers tend to use characteristics typically associated with the opposite gender while writing their reviews for example, if males use a majority of characteristics typically associated with females, and females use a majority of characteristics typically associated with males then the proposed method may become confused in predicting gender, as demonstrated in the error analysis section (Section 5.3).

6.2 Ethics

Our work aims to use the OpenAI GPT 3.5 model to perform a gender prediction task by incorporating custom prompting techniques. To ensure ethical considerations in our proposed method, we have only taken feedback from correctly classified reviews and not from misclassified ones during the feedback mechanism incorporation. We also ensure that the model is not hallucinating by conducting a detailed feedback analysis on the fine-tuned model.

Chapter 7: Conclusion and Future Work

7.1 Conclusion

In this thesis, we presented a novel approach that leverages Large Language Models (LLMs) for gender prediction from text without the need for extensive preprocessing. Our method incorporates custom prompts during both the fine-tuning and testing phases and adapts the testing prompt based on feedback from our model. This innovative feedback mechanism allows for continuous improvement and adaptation of the model. Our approach has shown promising results, achieving an average F1-score of 82%, which surpasses the performance of baseline methods. Additionally, we introduced a new dataset specifically curated for gender prediction from text and demonstrated the potential of modifying the testing prompt based on feedback from the fine-tuned model.

7.2 Future Work

Looking ahead, there are several opportunities for further research and development. The dataset used in this study was limited to anime reviews, and future work could explore the application of our method to other types of data to assess its generalizability. Expanding the scope of the dataset to include other gender-specific texts could provide deeper insights and enhance the model's accuracy. Moreover, investigating how to apply our approach to detect gender bias in text using LLMs could contribute significantly to the field of fair and ethical AI. By addressing these challenges, we aim to further refine our method and broaden its applicability in understanding and analyzing gender-related aspects in text data.

Chapter 8: Contributions

The contributions of this thesis are described in detail below –

(i) Due to the scarcity of datasets for this specific problem, we have created our own dataset, by minimizing the inclusion of AI-generated text.

(ii) Previous research has used machine learning or deep learning techniques to train models on text datasets to predict the author's gender. However, to our knowledge, this is the first study that has utilized a large language model to automatically identify gender-specific patterns in text data, achieving state-of-the-art performance.

(iii) This study used minimal preprocessing techniques on the input data and still achieved better results than previous studies that applied extensive preprocessing to the text data.

(iv) Previous research has primarily focused on interpreting models that perform well in predicting test data. However, with the advent of generative artificial intelligence, this research demonstrates how large language models (LLMs) can offer interpretations understandable to humans for incorrectly classified data inputs.

References

[1] Best practices for prompt engineering with openai api, 2021. Accessed: 2023-09-22.

[2] An introduction to large language models: Prompt engineering and p-tuning, 2021. Accessed: 2023-09-22.

[3] Fine-tuning gpt 3.5 with unstructured: A comprehensive guide, 2023. Accessed: 2023-09-22.

[4] Myanimelist.net, 2023. Accessed: 2023-05-01.

[5] Muhammad Abdul-Mageed, Chiyu Zhang, Arun Rajendran, AbdelRahim Elmadany, Michael Przystupa, and Lyle Ungar. Sentence-level bert and multi-task learning of age and gender in social media. *arXiv preprint arXiv:1911.00637*, 2019.

[6] James Briggs. Fine-tuning openai's gpt 3.5 turbo, 2023. Accessed: 2023-09-22.

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[8] Na Cheng, Rajarathnam Chandramouli, and KP Subbalakshmi. Author gender identification from text. *Digital investigation*, 8(1):78–88, 2011.

[9] Na Cheng, Rajarathnam Chandramouli, and KP Subbalakshmi. Author gender identification from text. *Digital investigation*, 8(1):78–88, 2011.

[10] V Sharmila Devi and S Kannimuthu. Author profiling in code-mixed whatsapp messages using stacked convolution networks and contextualized embedding based text augmentation. *Neural Processing Letters*, 55(1):589–614, 2023.

[11] Golnoosh Farnadi, Geetha Sitaraman, Shanu Sushmita, Fabio Celli, Michal Kosinski, David Stillwell, Sergio Davalos, Marie-Francine Moens, and Martine De Cock. Computational personality recognition in social media. *User modeling and user-adapted interaction*, 26:109–142, 2016.

[12] Dirk Hovy and Shannon L Spruit. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, 2016.

[13] Patrick Juola. Authorship attribution. *Foundations and Trends in Information Retrieval*, 1:233–334, 2007.

[14] Ksenia Lagutina, Nadezhda Lagutina, Elena Boychuk, Inna Vorontsova, Elena Shliakhtina, Olga Belyaeva, Ilya Paramonov, and P.G. Demidov. A survey on stylometric text features. In *2019 25th Conference of Open Innovations Association (FRUCT)*, pages 184–195, 2019.

[15] Sunakshi Mamgain, Rakesh C Balabantaray, and Ajit K Das. Author profiling: Prediction of gender and language variety from document. In *2019 International Conference on Information Technology (ICIT)*, pages 473–477. IEEE, 2019.

[16] Sunakshi Mamgain, Rakesh C Balabantaray, Ajit K Das, and Srikant Kumar. Author profiling: Predicting gender from document. In *Advances in Data Science and Management: Proceedings of ICDSM 2019*, pages 99–108. Springer, 2020.

[17] Reza Khan Mohammadi and Seyed Abolghasem Mirroshandel. Defeating author gender identification with text style transfer. *arXiv e-prints*, pages arXiv–2009, 2020.

[18] Dong Nguyen, Rilana Gravel, Dolf Trieschnigg, and Theo Meder. " how old do you think i am?" a study of language and age in twitter. In *Proceedings of the international AAAI conference on web and social media*, volume 7, pages 439–448, 2013.

[19] OpenAI. GPT-3.5 Turbo Updates. 2023. Accessed: 2023-09-25.

[20] Juan Pizarro. Profiling bots and fake news spreaders at pan'19 and pan'20: bots and gender profiling 2019, profiling fake news spreaders on twitter 2020. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 626–630. IEEE, 2020.

[21] Juan Pizarro. Profiling bots and fake news spreaders at pan'19 and pan'20: bots and gender profiling 2019, profiling fake news spreaders on twitter 2020. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 626–630. IEEE, 2020.

[22] Laria Reynolds and Kyle McDonell. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7, 2021.

[23] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1668–1678, 2019.

[24] Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. Social bias frames: Reasoning about social and power implications of language. *arXiv preprint arXiv:1911.03891*, 2019.

[25] H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791, 2013.

[26] Sina Sheikholeslami. Ablation programming for machine learning, 2019.

[27] Shaw Talebi. Fine-tuning large language models (llms). *Towards Data Science*, 2023. Accessed: 2023-12-02.

[28] Pradeep Vashisth and Kevin Meehan. Gender classification using twitter text data. In *2020 31st Irish Signals and Systems Conference (ISSC)*, pages 1–6. IEEE, 2020.

[29] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

[30] Kyle Wiggers. Openai brings fine-tuning to gpt-3.5 turbo. 2023. Accessed: 2023-09-22.

About the Author

Satya Uday Sanku is a Computer Science Master's student at the University of South Florida, Tampa, specializing in Natural Language Processing. He is passionate about making computers understand and use human language to facilitate better communication between people and machines. His thesis focuses on Author's Gender Classification using Natural Language Processing techniques and Large Language Models. Satya has contributed to the development of novel algorithms for gender detection and document retrieval. He has experience in various programming languages and machine learning libraries. Satya has also worked as a Graduate Teaching Assistant and an Android App Development Intern. His research has been published in reputable journals and conferences.