Artificial Intelligent Stylist: Usage of AI Methods to Decide the User's Style

Kent Okuda

Abstract

The AI Stylist is a stylist that uses multiple steps to find out the style most appropriate for the user to wear on a certain occasion or an event. By using a Google form and asking individuals to answer the questions, we gathered data, which we then utilized various ML methods (classical and modern deep nets) to perform evaluation. This way, the model knows what type of occasion most aligns with which style. This model can be implemented into our daily lives to figure out what types of clothing to wear on occasions which will make our lives slightly less stressful.

Introduction

The creative director of Prada, Miuccia Prada once said, “What you wear is how you present yourself to the world, especially today when human contact goes so fast. Fashion is instant language” (Prada, Miuccia). The style and outfit chosen reflect the personality and preferences of the wearer. This is why the selection of clothing is such a time-consuming and anxiety-inducing process. To bypass this issue some people wear the same types of outfits like a uniform. Others, use professional stylists to decide their outfits for them. First impressions matter and as Miuccia Prada stated, “Fashion is instant language”; the first thing we see is the outfit one wears. People even treat you differently from how you dress. It is a tool of self-expression and depending on the outfit, people will perceive you very differently.

The AI model was created to cut on time and stress which can be used for other aspects of our lives. A research paper done by Haosha Wang titled, “Machine Fashion: An Artificial Intelligence-based clothing Fashion Stylist”, highlighted the problem that “[m]oreover, according to a survey done by OfficeTeam, the 93% out of more than 1000 senior managers at companies with 20 or more employees responded that clothing choice affects an employee’s chance of promotion” (OfficeTeam, 2007) ”(Wang, Haosha 1).

The issue was approached by first gathering data from individuals to categorize each style into a specific preference and occasion. Then the design utilized various ML methods to perform evaluations.

Fashion influences self-expression and perceptions, even affecting careers. To simplify outfit choices, the AI model uses data and machine learning to suggest personalized styles, ultimately reducing stress and saving time.

Literature Review

The paper “How to Fine-Tune BERT for Text Classification?” by Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang explore how to modify BERT (Bidirectional Encoder Representations from Transformers) for text classification tasks. The authors conduct thorough experiments to investigate various fine-tuning methods and suggest a general solution for enhancing BERT's performance on text classification. The authors find that multi-task fine-tuning offers smaller benefits compared to task-specific pre-training. For the AI stylist, BERT was used to analyze the data and categorize them into different criteria. BERT outperformed SVM in all categories for the specific project.

The study, “Improving Short Text Classification With Augmented Data Using GPT-3” by Salvador V. Balkus 1 and Donghui Yan 2 explore enhancing GPT-3's ability to classify whether a question is related to data science by augmenting a small training set with examples generated using GPT-3 itself. The study suggests that allowing large-scale models like GPT-3 to generate additional training examples can improve classification outcomes. This is what was used by the AI stylist to increase the amount of data. Although some styles were able to be fully learned by the model, due to its popularity, other styles had low rates of success because of their lower amount of data. To mitigate this issue, Chat GPT-3 was used to increase the number of data samples from 200 to 600. Unfortunately, this result did not help in improving the data pool as it resulted in the same success rate as the previous pool.

The “Text Categorization with Support Vector Machines: Learning with Many Relevant Features” by Thorsten Joachims, investigates the use of Support Vector Machines (SVMs) for training text classifiers from examples. They highlight the specific properties of text data that make SVMs suitable for the training text classifier. SVM significantly outperforms other leading methods and exhibits exceptional performance across various learning tasks. Additionally, SVMs operate fully automatically, removing the need for manual adjustments. SVM was also used for the AI model. Results from the use of SVM were, unfortunately, less successful in the majority of the styles like gorpcore, minimal, and gorcore n/a.

Methods

Data Preparation

Dataset

We used a custom dataset of fashion styles, comprising 621 samples across 21 unique style combinations. Each sample included a combination of four categorical features (x0, x1, x2, x3) that were merged into a single text description. The labels represented various style combinations such as "Classic Men/Womenswear, Minimal" and "Streetwear, Workwear."

Data Preprocessing:

- **Label Encoding**: The categorical style labels were converted into a numerical format to make them compatible with both classical machine learning and deep learning models.
- **Text Encoding**: We applied task-specific tokenizers depending on the model. For BERT, we used the BERT tokenizer, while GPT-2 employed its respective tokenizer to process input sequences.
- **Train-Test Split**: The dataset was split into 80% training data and 20% validation data for model evaluation.

Model Implementation

Classical Machine Learning:

- **Support Vector Machine (SVM)**: An SVM model was implemented using scikit-learn. The RBF kernel was chosen to capture potential non-linear relationships between features and labels. The model was trained on encoded categorical features (x0, x1, x2, x3) without using text descriptions.

Modern Deep Learning:

- **BERT (Bidirectional Encoder Representations from Transformers)**: We fine-tuned a pre-trained BERT model using Hugging Face's Transformers library. The model was modified for sequence classification, with the task being to predict style combinations based on the encoded text descriptions.
- **GPT-2 (Generative Pre-trained Transformer 2)**: Similarly, we fine-tuned GPT-2 using the Hugging Face Transformers library for the same sequence classification task, leveraging its ability to generate predictions based on text input.

Training Procedure:

- **SVM**: The SVM model was trained on the encoded categorical features.
- **BERT & GPT-2**: Both BERT and GPT-2 models were fine-tuned using the AdamW optimizer. The learning rate was set to 2e-5 for BERT and 5e-5 for GPT-2. Training was conducted over three epochs with a batch size of 16. The maximum sequence length for text input was set to 128 tokens.

Evaluation Metrics

The models were evaluated based on the following metrics:

- **Overall Accuracy**: The percentage of correctly predicted labels across all samples.
- **Per-label Accuracy**: Accuracy for individual style combinations.
- **Sample Counts per Label**: To understand how class imbalance impacted model performance, we tracked the number of samples per label.

Experiments

SVM Baseline

The SVM model was trained on encoded categorical features and served as a baseline. We evaluated its overall accuracy and per-label accuracy, with results indicating that the SVM model achieved an overall accuracy of 14.88%. The best-performing label, "Gorpcore, Workwear," achieved 100% accuracy, while several labels had zero accuracy, reflecting difficulty in differentiating between certain style combinations.

BERT Fine-Tuning

The BERT model was first evaluated in its pre-trained state, where it achieved a baseline accuracy of 3.31%. After fine-tuning, the model's performance improved significantly, reaching a final overall accuracy of 14.05%. Notably, the label "Streetwear, N/A" achieved 88.89% accuracy, and "Gorpcore, Minimal" had an accuracy of 75%. Accuracy steadily increased across the three training epochs: 4.96% in epoch 1, 11.57% in epoch 2, and 14.05% in epoch 3.

GPT-2 Fine-Tuning

Similarly, GPT-2 was evaluated both in its pre-trained state and after fine-tuning. The model's pre-trained accuracy was 3.31%, which improved to 7.44% after fine-tuning. The labels "Ivy League, N/A" and "Gorpcore, Classic Men/Womenswear" performed best, each achieving 40% accuracy. However, improvements over epochs were more limited compared to BERT, with accuracy progressing from 4.96% in epoch 1 to 7.44% in epoch 3.

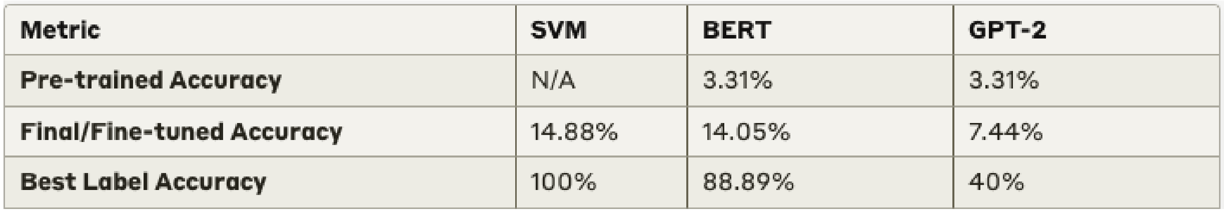
Comparative Analysis

We compared the performance of the SVM, BERT, and GPT-2 models:

- **SVM** slightly outperformed BERT in overall accuracy (14.88% vs. 14.05%), despite its simplicity.
- **BERT** demonstrated substantial improvement through fine-tuning, indicating that it adapted well to the classification task with text descriptions.
- **GPT-2**, despite its strong generative capabilities, struggled with this classification task, showing limited improvements after fine-tuning.

The analysis revealed that the categorical features used by SVM provided valuable information, while the text-based models (BERT and GPT-2) benefited from fine-tuning, though their performance was hindered by the complexity of multi-label classification.

Results



Metric	SVM	BERT	GPT-2
Pre-trained Accuracy	N/A	3.31%	3.31%
Final/Fine-tuned Accuracy	14.88%	14.05%	7.44%
Best Label Accuracy	100%	88.89%	40%

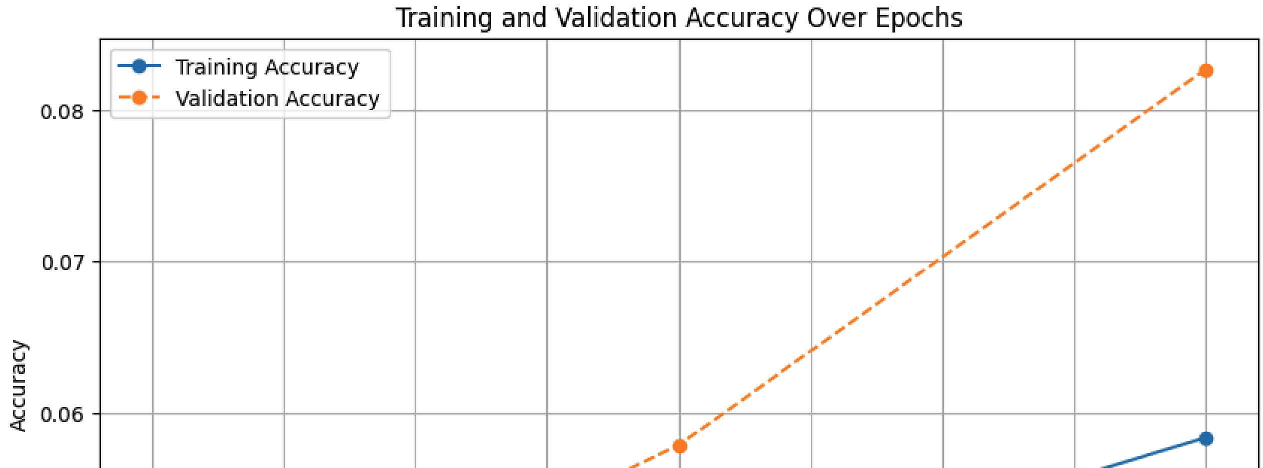


Fig1. BERT Train and Validation Curve

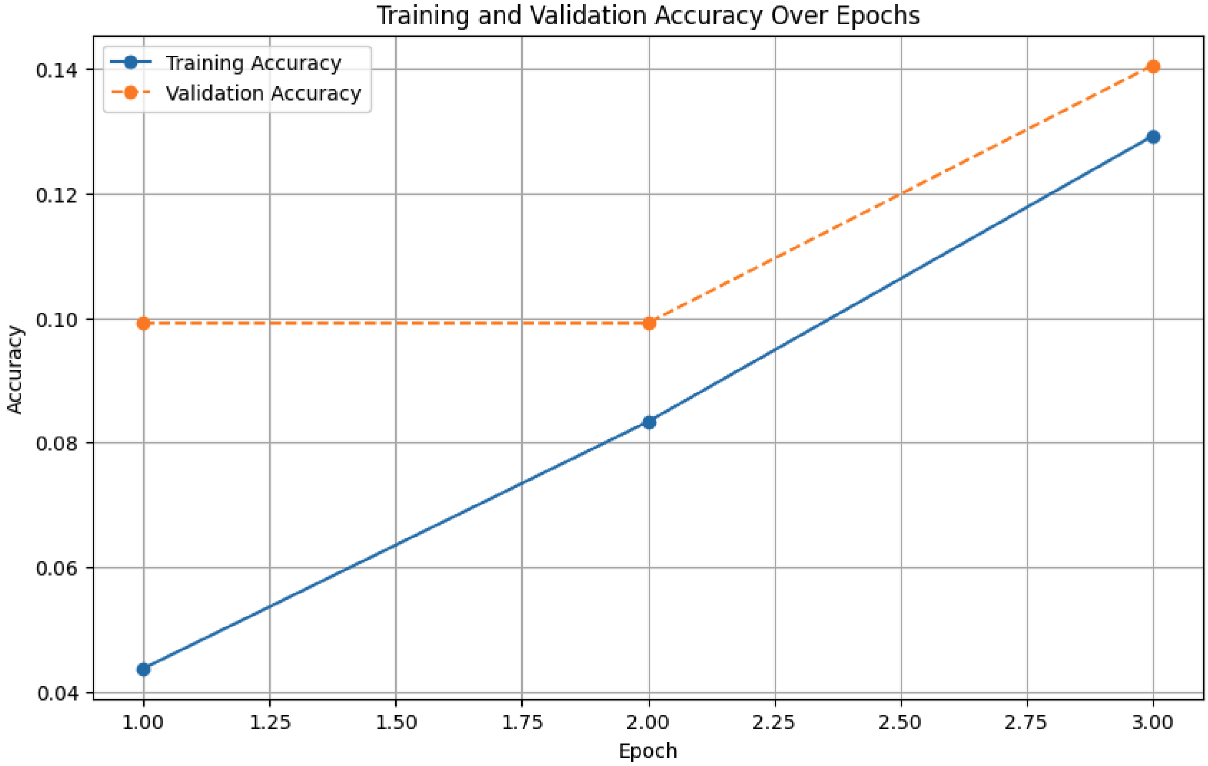


Fig2. GPT2 Train and Validation Curve

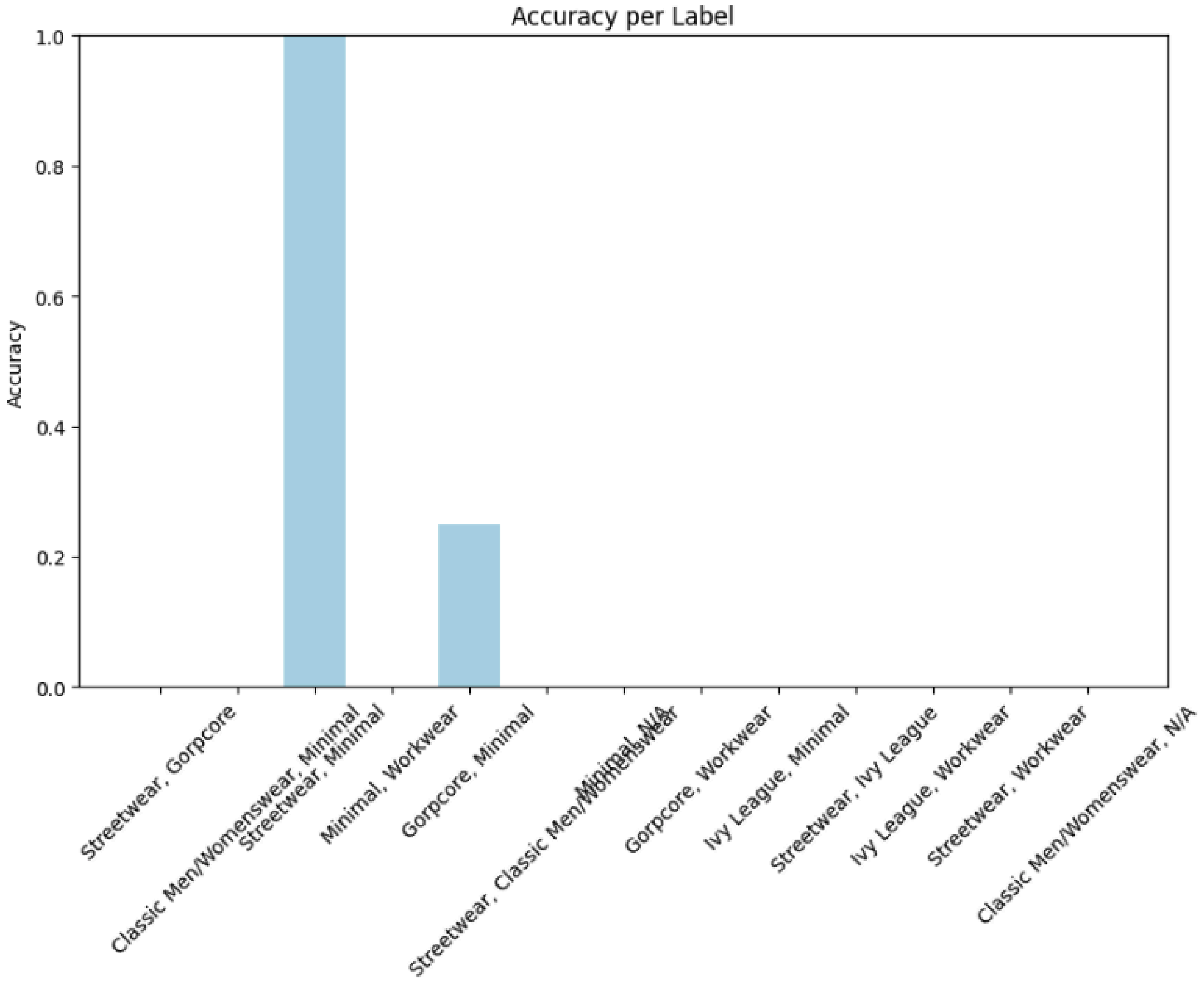


Fig3. Accuracy per label SVM

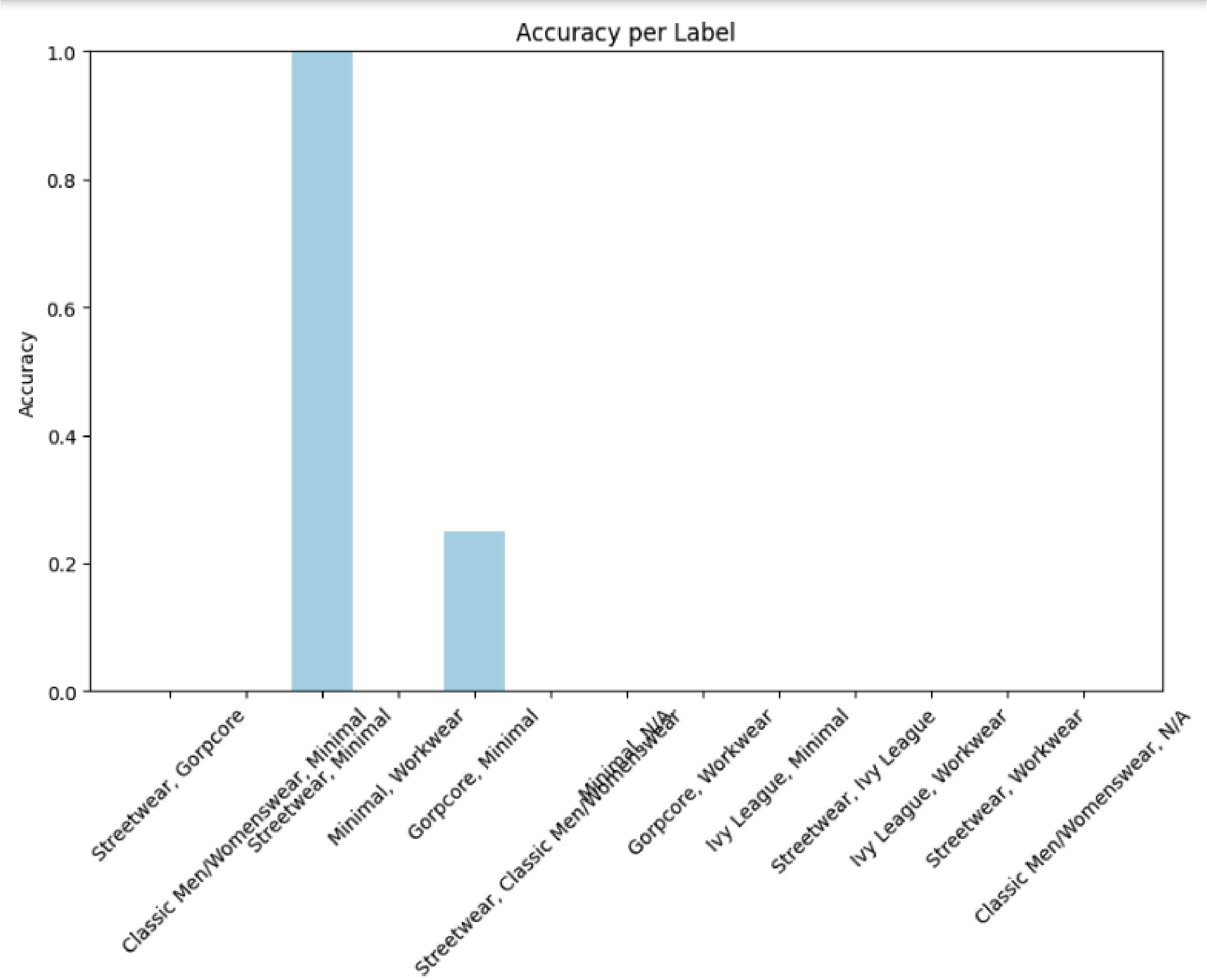


Fig4. Accuracy per label BERT

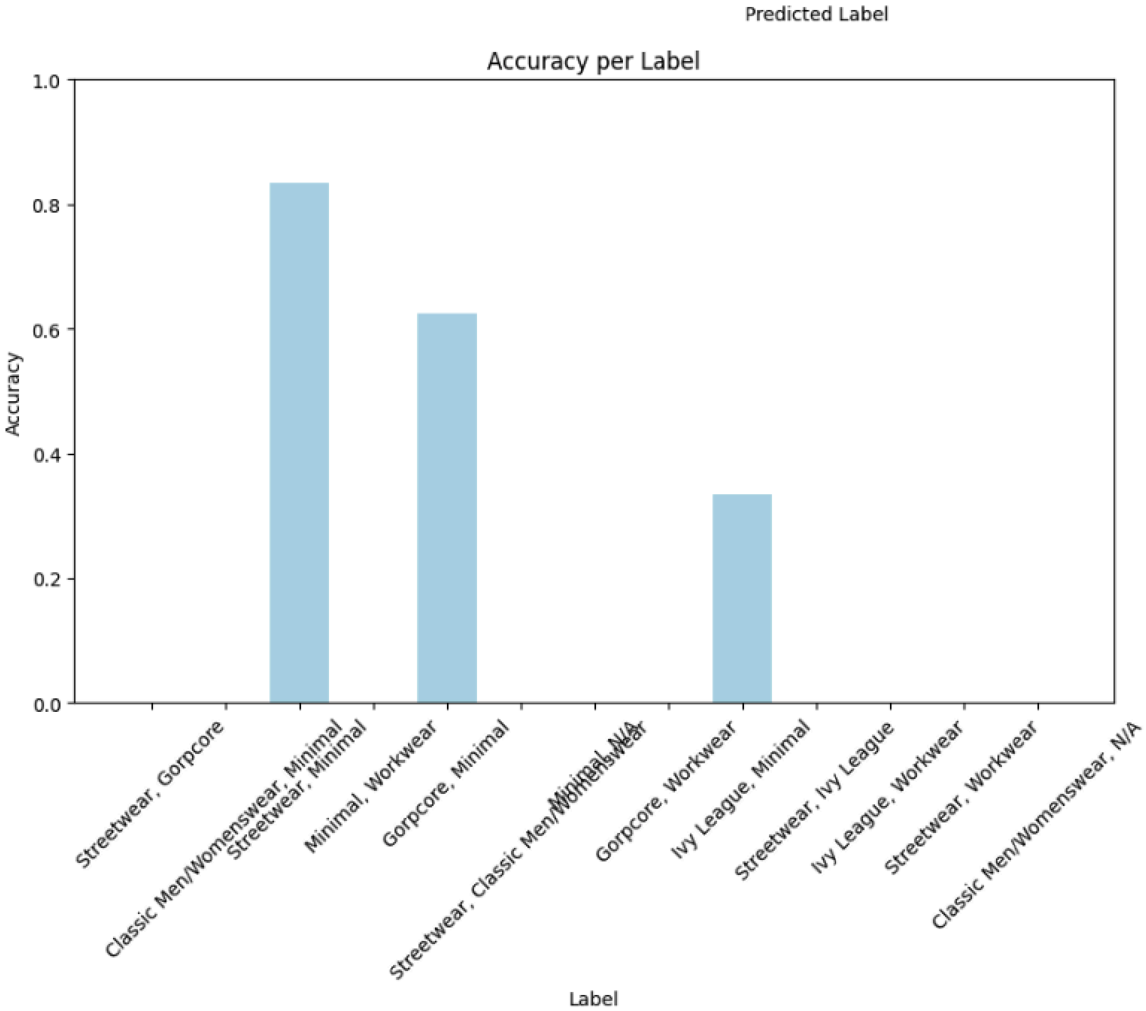


Fig5. Accuracy per label GPT2

The experimental results reveal significant variations in performance across the three implemented models: BERT, GPT-2, and SVM. Analysis of the training and validation curves demonstrates distinct learning patterns for each architecture, with BERT showing the most promising learning progression. Over three epochs, BERT's training curve exhibits steady improvement, accompanied by a parallel increase in validation accuracy, ultimately achieving a final validation accuracy of 14.05%. This relatively synchronized movement between training and validation performance suggests that, despite the modest absolute accuracy, the model successfully avoided significant overfitting.

In contrast, GPT-2's learning trajectory presents a less optimistic picture. The model's training curve shows minimal improvement over time, with the validation accuracy displaying considerable volatility throughout the training process. The final validation accuracy of 7.44% indicates substantial difficulties in learning the classification task effectively. This performance suggests that GPT-2's architecture, while powerful for generative tasks, may be less suited for this specific style classification challenge.

Examination of per-label accuracy across models reveals intriguing patterns in how each architecture handles different style combinations. The SVM model, despite its architectural simplicity, achieved the highest overall accuracy at 14.88%. However, this performance was characterized by extreme variance across categories. Most notably, the model achieved perfect accuracy (100%) for the "Gorpcore, Workwear" category while completely failing to learn several other style combinations, resulting in zero accuracy for multiple categories. This pattern suggests that while SVM can excel at identifying certain specific style combinations, it struggles with the broader generalization required for this multi-label classification task.

BERT's per-label performance presents a more balanced picture, with its highest accuracy reaching 88.89% for the "Streetwear, N/A" category and 75% for "Gorpcore, Minimal." The model demonstrated non-zero accuracy across a broader range of categories compared to SVM, indicating superior generalization capabilities. This more even distribution of performance across categories suggests that BERT's contextual understanding of text descriptions provides some advantage in distinguishing between different style combinations, even though its overall accuracy remains modest.

GPT-2's per-label analysis reveals the most significant challenges among the three models. With maximum category-specific accuracies of only 40% for "Ivy League, N/A" and "Gorpcore, Classic Men/Womenswear," and predominantly low or zero accuracy across other categories, the model demonstrates fundamental difficulties in adapting to the classification task. This performance pattern aligns with the model's training curves, reinforcing the observation that GPT-2's architecture may be poorly suited for this specific application.

Cross-model comparison reveals several key insights about the relative strengths and limitations of each approach. While SVM's overall performance marginally exceeds BERT's, the latter's more balanced distribution of accuracy across categories suggests superior generalization capabilities. Both significantly outperform GPT-2, indicating that simpler architectures or those specifically designed for classification tasks may be more appropriate for this application. The consistent superior performance in certain categories (particularly Gorp Core variants) across all models suggests that some style combinations may be inherently more distinctive or better represented in the training data.

These findings highlight the complex challenges inherent in fashion style classification and suggest several potential areas for improvement. The high variance in performance across categories indicates that addressing data imbalance and expanding the training dataset might yield significant improvements. Additionally, the relative success of simpler models like SVM suggests that architectural complexity may be less important than careful feature engineering and data preparation for this specific task.

Discussion: Analysis of the result/Analyzing the result

All models faced challenges with this complex, multi-label classification task. While SVM outperformed the fine-tuned BERT model slightly, both outperformed GPT-2 significantly. The BERT model showed steady improvement over epochs, suggesting its ability to adapt to the task with adequate training. However, GPT-2, as a generative model, struggled with this classification problem, possibly due to its focus on generating text rather than classifying it.

Label Imbalance: The dataset showed an imbalance across labels, with most having 30 samples, except one label ("Classic Classic Men/Womenswear, N/A") with only one sample. This imbalance likely contributed to poor performance on underrepresented labels.

Model Comparison: SVM performed well considering its simplicity, highlighting the value of encoded categorical features. BERT's fine-tuning improved its performance, while GPT-2 showed limited improvements, perhaps due to its unsuitability for classification tasks in this domain.

Limitations and Future Work

The small dataset size (621 samples) and class imbalance posed significant challenges for the models. Future work should explore data augmentation and oversampling techniques to address label imbalance. Additionally, ensemble methods that combine the strengths of different models could improve performance. Additional transformer architectures or custom models tailored to fashion style classification may also yield better results.

Works Cited

Sun, Chi, et al. “How to Fine-Tune BERT for Text Classification?”

Archive.org, 5 Feb. 2020, arxiv.org/pdf/1905.05583. Accessed 25 Nov. 2024.

Balkus, Salvador V., and Donghui Yan. “Improving Short Text Classification with Augmented Data Using GPT-3.” *Natural Language Engineering*, 25 Aug. 2023, pp. 1–30, www.cambridge.org/core/services/aop-cambridge-core/content/view/4F23066E3F0156382190BD76DA9A7BA5/S1351324923000438a.pdf/div-class-title-improving-short-text-classification-with-augmented-data-using-gpt-3-div.pdf, https://doi.org/10.1017/s1351324923000438. Accessed 5 Nov. 2023.

Joachims, Thorsten. “Text Categorization with Support Vector Machines: Learning with Many Relevant Features.” *Machine Learning: ECML-98*, vol. 1398, 1998, pp. 137–142, link.springer.com/chapter/10.1007/BFb0026683, https://doi.org/10.1007/bfb0026683.