CS552J Assessment 2 - To what extent do cross-generational dialogues align with youth-like or old-like conversation styles?

Tharushi Suwaris
Department of Computing Science
University of Aberdeen
52426304
t06ts24@abdn.ac.uk

1 Introduction

This study investigates how stylistic features in dialogue vary across generations and whether these can be used to classify conversations as youth-like or old-like. Using a dataset of age-labeled dialogues, the features are extracted such as average utterance length, swear word usage, filler word frequency, vocabulary overlap, and lexical repetition. Two machine learning models, Logistic Regression and XGBoost, are trained on the Young-Young (Y-Y) and Old-Old (O-O) dialogues to perform binary classification. These models are then applied to cross-generational (Old-Young (O-Y)) dialogues to examine how they are categorized, and which features influence the decisions. SHAP analysis is employed to interpret model outputs and identify the most impactful linguistic parameters. The findings show that swear word frequency plays a key role in classification, while filler words and vocabulary usage also contribute. This research highlights how machine learning can model age-related language variation and provide insight into cross-generational communication styles.

2 Description of Data and Methods

2.1 Data

The dataset used in this project consists of conversational transcripts sourced from the Spoken British National Corpus 2014 (Spoken BNC2014[1]). Each dialogue consists of multiple turns between two or more speakers, and includes rich metadata such as speaker age, gender, and nationality, which were from volunteer participants across the United Kingdom.

To address the problem of "To what extent do cross-generational dialogues align with youth-like or old-like conversation styles?", the dataset is processed in such a way that, it has only two speaker dialogues, and the age categories were selected as 19-29 as young and 50 above as old, making the age gap between the groups are 20 years apart, following [Lennert Jansen and Pezzelle, 2021] . The number of dialogues in these age categories can be seen as in Figure 1.

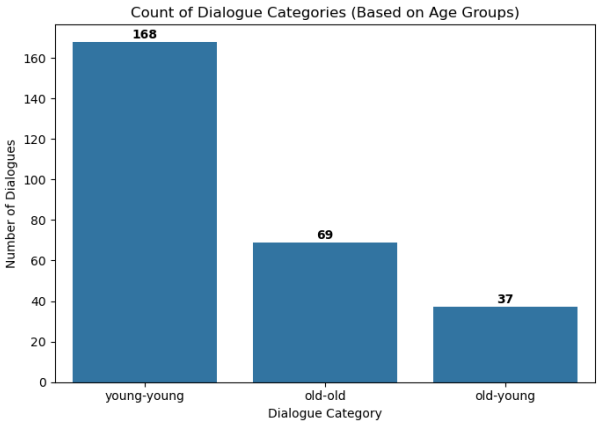


Figure 1: Count of dialogue categories based on the age categories.

To explore the dialogue data, several linguistic and interactional parameters were analyzed. These include average utterance length, the unique word ratio (calculated both per dialogue and per speaker), utterance proportion per speaker, the frequency of swear words and filler words. The list of swear words was drawn from the terms most frequently used in the BNC2014 corpus, as identified by [Love, 2021]. For filler words, non-lexical vocalizations indicating hesitation, such as "uh," "um," and similar expressions, were included, following [Huang., 2024]. Further details on how these features were computed are provided in the appendix A.

In addition to these manually derived features, the Dialign[2] tool is used to extract higher-level interactional metrics. These include:

[1] https://cass.lancs.ac.uk/cass-projects/spoken-bnc2014/

[2] https://github.com/GuillaumeDD/dialign

- Vocabulary Overlap: the degree of shared vocabulary between speakerse,
- Expression Variety (EV): the diversity of expression instances,
- Expression Repetition (ER): the repetition frequency of shared expressions,
- ENTR: a measure of the lexical entropy of shared expressions, reflecting their complexity,
- L: the average length in tokens of shared expression instances.

These parameters were analyzed across different categories of age groups to identify stylistic and interactional differences. According to Figure 10, the parameters except average filler count and overall unique word count follow a similar pattern, where the Y-Y category has the highest value followed by O-Y category, leading the O-O category to the lowest value. These value differences are not significant except for the average swear count. Regarding the parameters average filler count and overall unique word count also follow a similar pattern, where the older group tends to have the highest value, followed by the O-Y category. These patterns suggest that stylistic markers like swearing, and lexical variety vary systematically by age, providing useful cues for classification.

Additionally, a one-way ANOVA[3] was conducted to assess the statistical significance of each parameter across the age categories. This test enabled the comparison of mean values for features between the different age groups. As presented in Table 1, most parameters demonstrated statistically significant differences ($p < 0.05$) across age categories, with the swear word count showing the highest level of significance.

In contrast, average utterance length yielded an F-statistic of 1.096 and a p-value of 0.335, indicating no statistically significant difference across age groups. Since this p-value exceeds the conventional threshold of 0.05, suggesting that the group means are not significantly different. Based on these results, all features except average utterance length were retained for further analysis and model training in the dialogue classification task.

[3] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.f_oneway.html

Feature	F-value	p-value
Average utterance length	1.096	0.335
Swear word count	47.110	0
Filler word count	21.153	0
Unique words ratio per dialogue	4.810	0.0085
Expression Variety	28.768	0
Expression Repetition	11.290	0
Vocabulary Overlap	43.130	0
ENTR	16.990	0
L	21.024	0

Table 1: ANOVA results for each linguistic feature across dialogue categories.

2.2 Models/Algorithms

To explore how dialogue features influence the classification of cross-generational conversations as either youth-like or older-like, two classification models were employed: Logistic regression[4] and XGBoost[5]. The approach taken to identify how the models identify cross generational dialogues as more youth like or old like, it is first trained with Y-Y and O-O categories, as a allowing the models to behave like a binary classification model, then separately find out how the trained models categorize actual O-Y category as either Y-Y or O-O, and the parameters which contributes to these decision are observed.

To observe the desired observations, logistic regression model is used considering its efficiency and fast nature. This model is also known to be interpretable, meaning it considers the feature importance. Also, the XGBoost model, a gradient-boosted decision tree ensemble, was selected for its robustness and strong predictive performance, particularly in handling complex feature interactions.

To further interpret the model decisions, especially for the O-Y category classifications, SHAP (SHapley Additive exPlanations)[6] analysis was used. SHAP values provide a unified measure of feature contribution for individual predictions, enabling us to understand which features most strongly influenced the model's decision for each dialogue.

[4] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

[5] https://xgboost.readthedocs.io/en/release_3.0.0/

[6] https://shap.readthedocs.io/en/latest/

2.3 Experimental Approach

The experimental setup includes several key preprocessing and evaluation steps to ensure fair assessment of the model.

Preprocessing: All input features were standardized using the StandardScaler[7] from Python's scikit-learn library to normalize feature scales. No rebalancing techniques were applied, and the original class distribution was preserved to avoid losing important features.

Evaluation Metrics: Given the class imbalance, multiple evaluation metrics were used, including precision, recall, F1-score, and balanced accuracy[8]. The balanced accuracy metric was particularly important as it accounts for imbalanced class distributions by averaging recall across classes, reducing bias toward the majority class.

Cross-validation: To assess model generalization, 5-fold cross-validation was conducted on the training data, which consisted of 80% of the full dataset. The remaining 20% was held out as a final test set. This setup ensures robust performance estimation.

3 Results

3.1 Cross-Validation Results

To evaluate the performance and generalizability of the model, 5-fold cross-validation was applied on the training set. The average results across the folds for each evaluation metric are detailed in Table 2.

Metric	LR	XGBoost
Precision	0.694	0.749
Recall	0.881	0.81
F1 Score	0.774	0.77
Balanced Accuracy	0.848	0.841

Table 2: Average cross-validation scores for models (LR- Logistict Regression and XGBoost) across all evaluation metrics.

Both models demonstrate strong performance, with the XGBoost model slightly outperforming Logistic Regression in terms of precision, while Logistic Regression shows higher recall. This indicates that Logistic Regression was more sensitive to the positive class, whereas XGBoost made fewer false positives. The F1 scores and balanced accuracy metrics are relatively similar across both models, suggesting similar overall performance.

3.2 Final Test Results

The final evaluation was conducted on the held-out test set (20% of the dataset). The results are summarized in Table 3.

The Logistic Regression model showed moderate recall (0.8000) on the test set, but its precision (0.4000) and F1 score (0.5333) decreased notably, indicating an increase in false positives (as indicated in Figure 2) and reduced reliability in its predictions. The XGBoost model, while also experiencing a drop in performance compared to validation, maintained better precision (0.5000) and recall (0.8000) balance, resulting in a slightly higher F1 score. As shown in Figure 3, XGBoost made fewer false positives than Logistic Regression, suggesting better generalization and more reliable class discrimination.

Metric	LR	XGBoost
Precision	0.4	0.5
Recall	0.8	0.8
F1 Score	0.53	0.6
Balanced Accuracy	0.74	0.79

Table 3: Final test set performance for both models.

3.3 Cross-Generational Dialogue Classification

To assess how the models classify cross-generational (O-Y) dialogues, the above trained models are used. This approach allowed to observe whether the dialogue style in these mixed-age interactions leaned more towards a younger or older communication pattern.

Table 5 summarizes the distribution of predicted labels for the cross-generational dialogue instances.

Model	Y - Y	O - O
Logistic Regression	20	17
XGBoost	21	16

Table 4: Predicted classifications of cross-generational dialogues by each model.

These results indicate that both models slightly favored classifying cross-generational conversations as more youth-like in style, though the difference between the two categories was relatively

[7] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

[8] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.balanced_accuracy_score.html

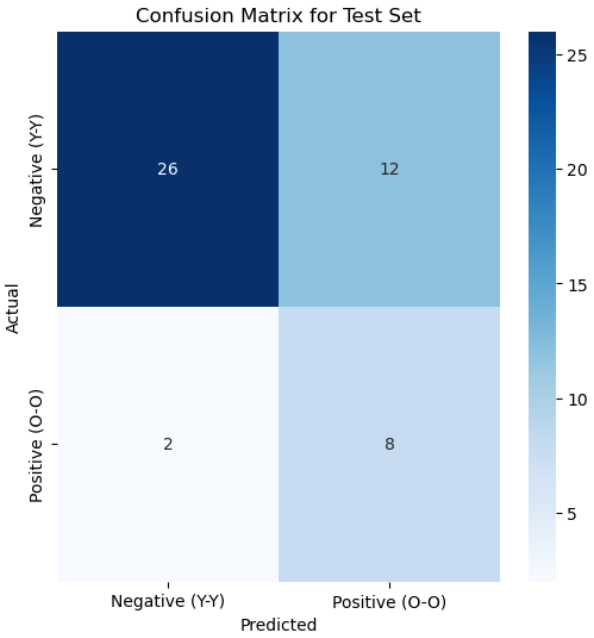


Figure 2: Confusion matrix for logistic regression test data

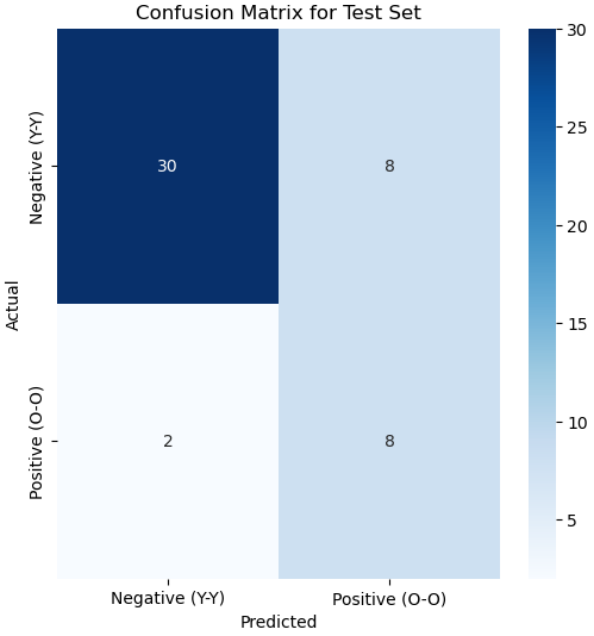


Figure 3: Confusion matrix for XGBoost test data

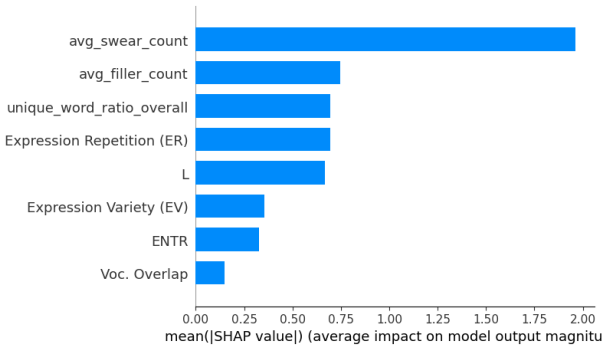


Figure 4: Feature importance -logistic regression

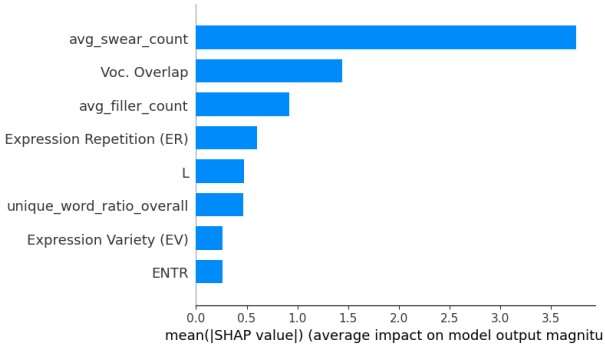


Figure 5: Feature importance -XGBoost

small. This pattern suggests that mixed-age dialogues may exhibit stylistic blending, with a slight tendency toward the linguistic patterns of younger speakers.

3.4 SHAP Analysis

The SHAP analysis (Figures 4 and 5) was conducted to interpret the contribution of each feature to the model predictions. Across both models, the average swear count emerged as the most influential parameter, showing a significantly higher mean SHAP value compared to all other features. This dominance is visually evident in the SHAP summary plots, where there is a clear drop-off in importance after this feature.

For the XGBoost classifier, vocabulary overlap was identified as the second most impactful feature. In contrast, for the logistic regression model, vocabulary overlap was the least significant feature, indicating differing sensitivities of the models to this parameter. Another feature that consistently demonstrated a meaningful impact across both models was the average filler count.

Figure 9 provides a more focused view by comparing SHAP values for cross-generational (O-Y) instances, depending on whether they were predicted as youth-like (Y-Y) or old-like (O-O). In

both models, the average swear count remains the top driver of prediction, regardless of classification outcome.

For the XGBoost model, the top three features influencing both Y-like and O-like predictions remained consistent; average word count, vocabulary overlap, and average filler count in order. In contrast, the logistic regression model exhibited different feature influences for the two predicted classes. For youth like predictions, the top contributing features were average swear count, overall unique word count ratio and expression repetition, while for the old-like predictions, the average swear count, average filler count and overall unique word count ratio are the most significant.

4 Discussion & Conclusion

This study explored how stylistic features influence the classification of cross-generational dialogues as youth-like or old-like. Logistic regression and XGBoost were used. As observed in Figure 6 in appendix, the fact the greatest number of swear words are used in the age category of 19-29 and other older generations used comparatively less must have created the most impact when training the model. The SHAP analysis further reveals that while both models heavily rely on features such as swear count and filler words (used here as hesitation markers). Also, it seems like the XGBoost model has learned some complex interactions between features such as vocabulary overlap given its non-linearity feature. Overall, both models primarily learned from basic yet expressive linguistic markers that distinguish generational dialogue styles.

Cross-generational dialogues were classified almost equally into youth-like and old-like categories, reflecting stylistic blending. This shows that simple features can meaningfully represent generational style. Future research could benefit from more complex models and additional features such as syntax or semantic structure.

References

Qiqi Huang. 2024. A corpus-based comparative analysis on the use of non-lexical filler words in spoken english between chinese efl learners and native english speakers. In *Journal of Humanities, Arts and Social Science*, pages 266–272.

Margot J. van der Goot Raquel Fernández Lennert Jansen, Arabella Sinclair and Sandro Pezzelle. 2021. Detecting age-related linguistic patterns in dialogue: Toward adaptive conversational systems. *Proceedings of the Eighth Italian Conference on Computational Linguistics CliC-it*.

Robbie Love. 2021. Swearing in informal spoken english: 1990s–2010s. *Text Talk*, 41(5-6):739–762.

A Exploring Utterance Data

The following parameters were extracted to observe the utterance data.

- Total words of dialogue
- Total words per speaker (per dialogue)
- Unique words per dialogue
- Unique words per speaker (per dialogue)
- Number of utterances (per dialogue)
- Utterance counts per speaker – number of turns taken by a particular speaker in a dialogue
- Overall unique word ratio = unique words per speaker / total words per speaker
- Unique word ratio per speaker = unique words per speaker / unique words per dialogue
- Utterance proportion per speaker = utterance count per speaker / number of utterances
- Average swears count (per dialogue per speaker)
- Average filler words count (per dialogue per speaker)

Figure 6 presents how average swear word count and average filler word count vary by age group. Swear word usage is most prominent among younger age groups (11–18, 19–29), with the highest frequency in the 19–29 category, and declines with older groups. In contrast, filler words (hesitation markers) are more prevalent in older age categories, with the 90–99 age group showing the highest usage, and younger speakers using them the least.

Figure 7 shows how overall unique word ration and unique words ratio per speaker varies depending on the age categories. Figure 8 shows how Overall unique word ration and unique words ratio per speakers vary depending on the number of speakers in a dialogue, the unique word ration per

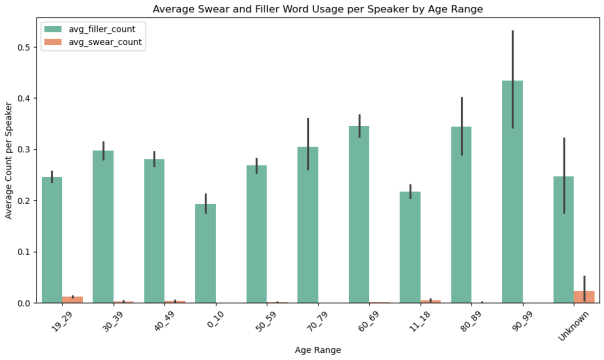


Figure 6: Average filler counts and swear word count variation depending on the age categories.

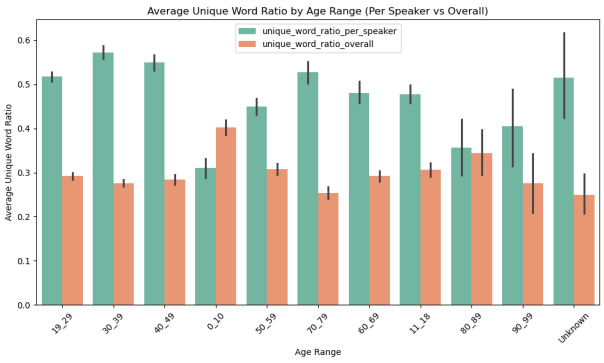


Figure 7: Overall unique word ration and unique words ratio per speaker variation depending on the age categories.

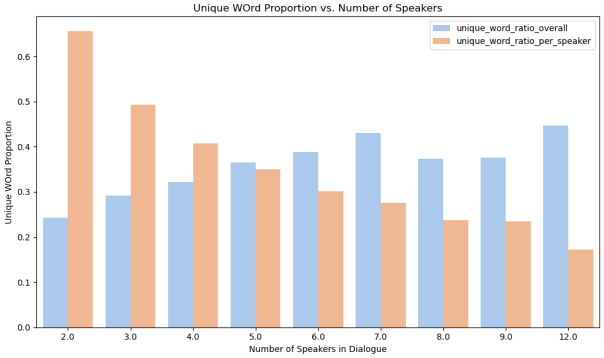
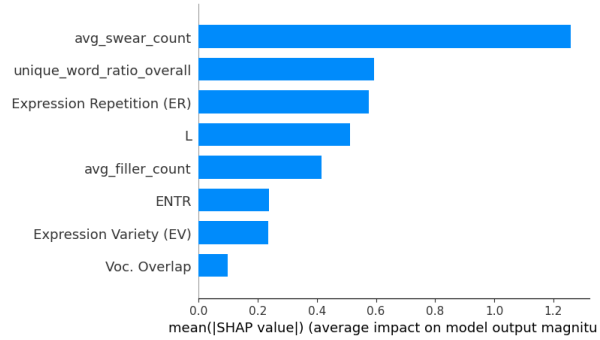


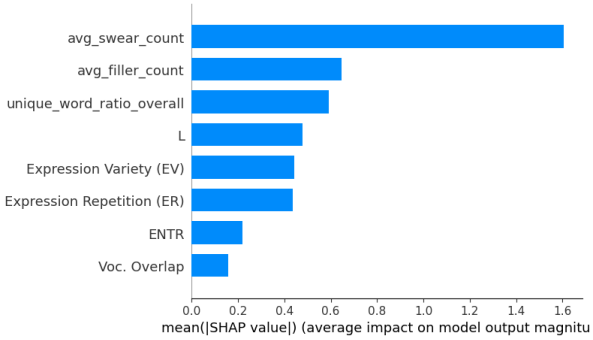
Figure 8: Overall unique word ration and unique words ratio per speaker variation depending on the number of participants.

speaker gradually decreases with the number of participants having the highest value for two speaker participants while the overall unique word ratio shows a gradual increase with some fluctuations with the number of participants.

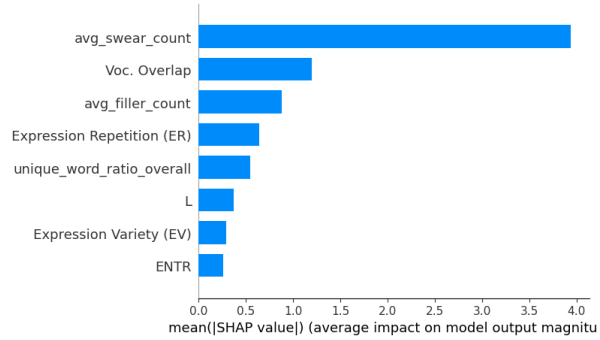
The report contains 2006 words excluding references and appendix.



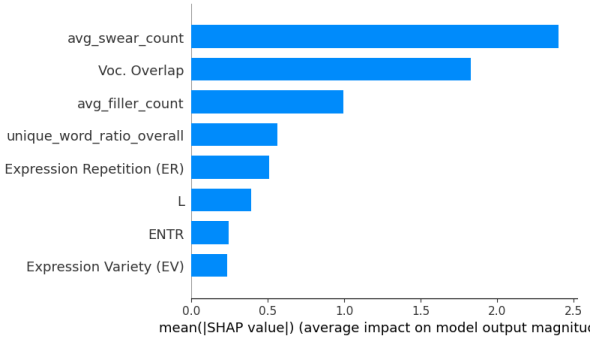
(a) SHAP summary for O-Y instances predicted as Y-Y for logistic regression



(b) SHAP summary for O-Y instances predicted as O-O for logistic regression

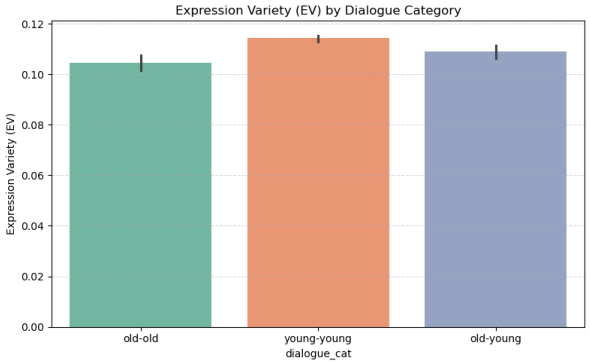


(c) SHAP summary for O-Y instances predicted as Y-Y for XGBoost

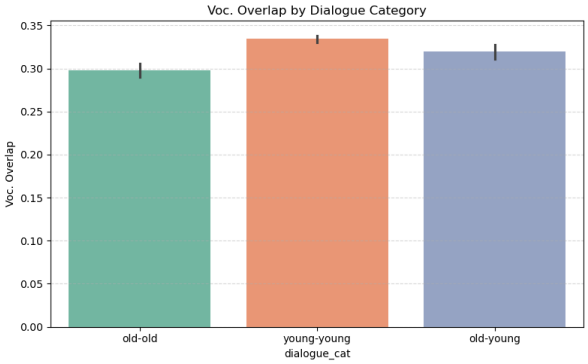


(d) SHAP summary for O-Y instances predicted as O-O for XGBoost

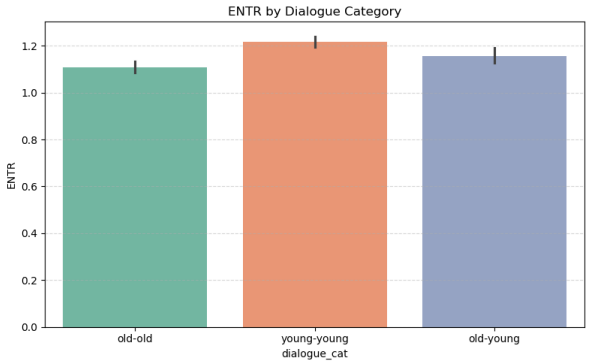
Figure 9: SHAP summary for O-Y instances predicted as Y-Y and O-Y instances predicted as O-O



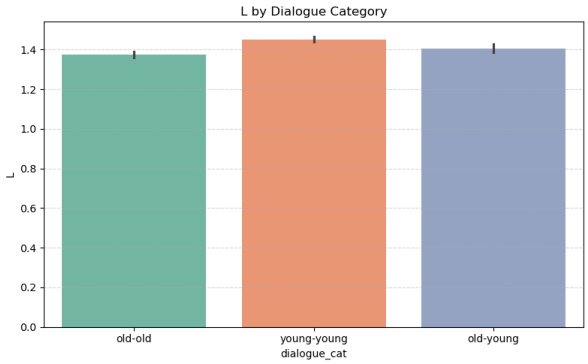
(a) Mean expression variety by dialogue categories



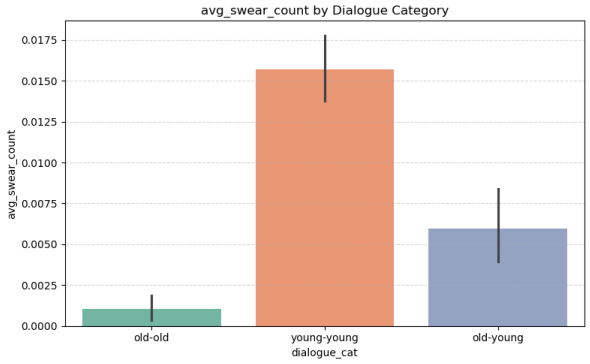
(b) Mean vocabulary overlap by dialogue categories



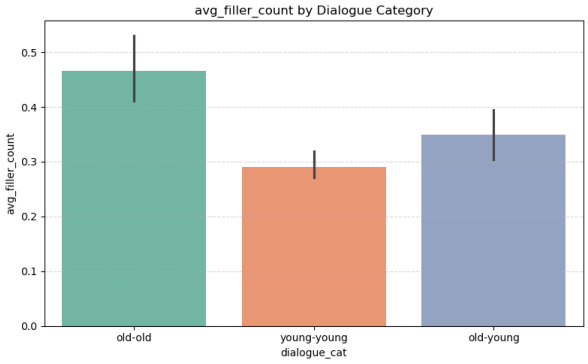
(c) Mean ENTR by dialogue categories



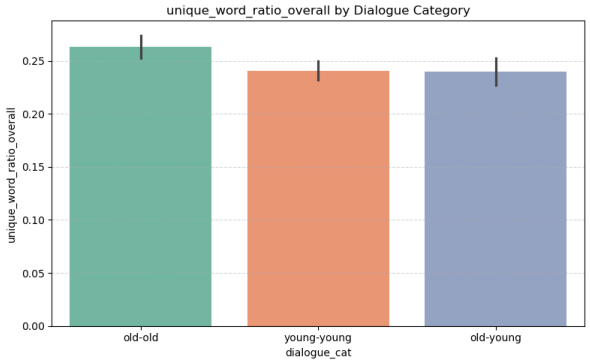
(d) Mean L by dialogue categories



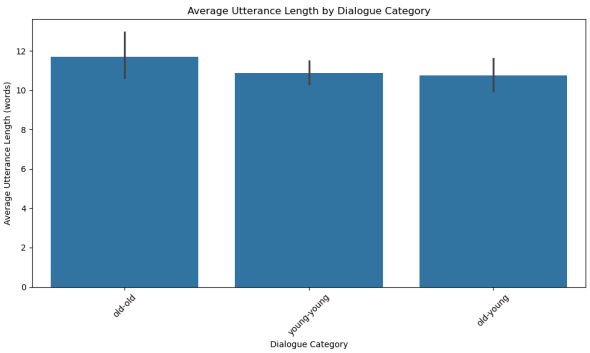
(e) Average swear words by dialogue categories



(f) Average filler words by dialogue categories



(g) Mean unique word ratio per dialogue by dialogue categories



(h) Mean utterance length per dialogue by dialogue categories

Figure 10: Feature Distributions by Dialogue Categories