Efficiency of Class Naming at MIT—Numbers or Names?

Matthew McManus
Massachusetts Institute of Technology
Department of Electrical Engineering and Computer Science
mattmcm@mit.edu
and
Daniel Kim
Massachusetts Institute of Technology
Department of Electrical Engineering and Computer Science
dyk0518@mit.edu

Abstract

Past studies have shown that different naming schemes across the globe have evolved to reach near-maximum optimality in terms of their trade-offs between accuracy and efficiency. However, it is not clear if this linguistic optimality is also present in smaller, more local environments. This paper aims to answer this question by investigating the efficiency of language naming conventions for classes at the Massachusetts Institute of Technology (MIT). We analyze factors influencing how classes are referred to by students and faculty, considering both formal and informal naming contexts. Our findings contribute to the understanding of language evolution and efficiency in specialized academic domains.

1 Introduction

At MIT, every class is assigned a unique number that efficiently compresses information about the course. This naming system within MIT allows students to then refer to their classes by name and number. For example, if you were to say I am taking "6.009" or "double-oh-nine," most people know that you are taking a class called "Fundamentals of Programming".

The unique class naming scheme employed by MIT prompts questions regarding its efficacy in information communication. While previous research has explored how various naming systems around the world tend to evolve to attain an optimal balance between complexity and accuracy, it remains uncertain whether this linguistic optimality extends to smaller academic environments. Therefore, this study aims to assess the efficiency of class naming conventions at MIT, investigating factors influencing how classes are referenced in both formal and informal campus discourse. Additionally, we examine variables such as course level, department affiliation, and official name length, which may impact preferred naming conventions. Furthermore, our analysis reveals instances of linguistic efficiency through the utilization of abbreviations, nicknames, and other shorthand references.

In this study, we employ a mixed-methods approach that combines survey data, computational analysis, and data visualization techniques. By surveying a diverse sample of MIT students and analyzing the collected data using logistic regression and neural network models, we aim to uncover the underlying patterns and relationships between class attributes and naming conventions. Our study can bridge the gap between theoretical principles of language evolution and their practical applications in real-world settings. Through this interdisciplinary approach, we contribute to the ongoing discourse on language evolution and efficiency, while also offering practical insights that can inform communication practices within academic institutions. Specifically, we extend the application of information-theoretic principles to the domain of class naming conventions at MIT.

2 Related Work

Language evolution and efficiency have been central themes in linguistic research, providing valuable insights into how languages adapt to the cognitive and communicative needs of their users. The principle of least effort, introduced by Zipf (Zipf, 1949) and further developed by Ferrer i Cancho & Solé (2003) (Ferrer i Cancho and Solé, 2003), suggests that languages evolve to minimize the articulatory and cognitive effort required for communication. This principle manifests in various linguistic phenomena, such as the abbreviation of common words and the preference for shorter, more frequent terms.

In recent years, the concept of communication efficiency has gained prominence in the study of language evolution, particularly in the context of

color naming systems. Zaslavsky et al. (Zaslavsky et al., 2018) demonstrated that color-naming systems across languages achieve near-optimal compression by optimizing the information bottleneck (IB) trade-off between the complexity and accuracy of the lexicon. They found that small changes in a single trade-off parameter account for much of the observed cross-language variation in color naming. Additionally, they showed that efficient IB color-naming systems exhibit soft category boundaries and often leave large regions of color space inconsistently named, phenomena that are also found empirically.

Building upon this work, Zaslavsky et al. (Zaslavsky et al., 2022) further explored the evolution of color naming systems, providing evidence that the evolution of color terms reflects a pressure for efficiency. By analyzing color naming data from the recent past, they demonstrated that the IB principle can account for the historical trajectory of color term evolution, capturing both discrete and continuous aspects of this process.

The emergence and cultural evolution of linguistic structures have also been investigated through experimental studies. Kirby et al. (Kirby et al., 2008) conducted laboratory experiments to observe how language evolves through repeated interactions among individuals, providing insights into the mechanisms that drive the formation and adaptation of linguistic conventions. These findings underscore the importance of social learning and cultural transmission in shaping language evolution.

While these studies have significantly advanced our understanding of language evolution and efficiency, there remain gaps in their application to specialized domains. Technical fields and academic settings, such as MIT, present unique linguistic environments where efficiency pressures may manifest differently than in general language use. Ellis (Ellis, 2008) explores the dynamics of second language acquisition, emphasizing the role of frequency, salience, and contingency in shaping language learning and usage patterns. This work highlights the potential for distinct evolutionary dynamics in specialized contexts, where the demands of the domain may influence the formation and adoption of linguistic conventions.

Drawing inspiration from Zaslavsky et al.'s work on efficient compression in color naming systems, we aim to extend the application of information-theoretic principles to the domain of class naming conventions at MIT. By adapting their methodology and insights to our specific context, we seek to uncover similar patterns of efficient compression and its role in shaping the evolution of naming conventions within this specialized academic domain. Our study builds upon the theoretical foundations laid by previous research while leveraging novel data sources and analytical methods to provide a more comprehensive understanding of how efficiency pressures shape linguistic conventions in this unique setting.

3 Data and Methods

In this section, we introduce the data collection process, as well as the analysis of the data collected. Then, we walk through how these data were processed and analyzed by different methods, including logistic regression and neural network models.

3.1 Data Collection

To investigate the efficiency of language naming conventions for classes at MIT, we conducted a survey of current MIT students to gather information about their class naming practices.

3.1.1 Survey Design and Participants

We designed an online survey to elicit responses from MIT students regarding how they verbally refer to their classes. The survey included the following questions:

- Year (e.g., Freshman, Sophomore, Junior, Senior, Graduate)
- Official name of classes enrolled in during Fall 2023 and Spring 2024 semesters
- Preferred way of referring to each class (name or number)

The survey was distributed to MIT students via dorm-spam (an email list that includes most of the entire undergraduate population) which helped target a diverse range of participants across different departments and years of study.

3.1.2 Survey Results

A total of 116 MIT students participated in the survey, providing data on 350 classes. Figure 1 shows the distribution of participants by year: 36 Freshmen, 22 Sophomores, 28 Juniors, 23 Seniors, and 7 Graduate students. As seen in Figure 2, the most represented departments in the survey were

Course 6 (Electrical Engineering and Computer Science), Course 18 (Mathematics), and Course 21M (Music and Theater Arts). The survey results

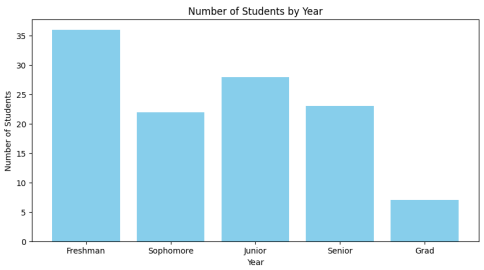

Figure 1: Distribution of participants by year.

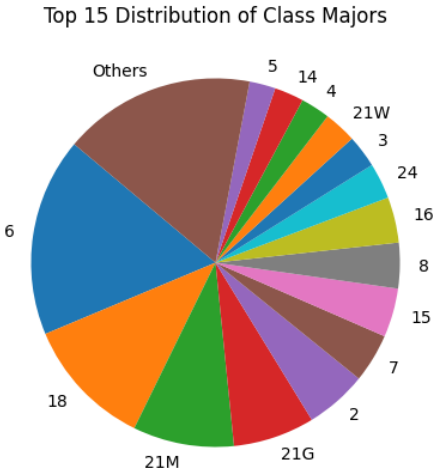

Figure 2: Most represented departments in the survey.

revealed that 53.71% of classes were referred to by name, while 46.29% were referred to by number. Among classes referred to by name, 38.83% used the full official name, and 61.17% used a shortened or nickname version. For classes referred to by number, 15.43% used the old course number, and 84.57% used the new or current number.

The survey results provide valuable insights into the overall distribution of class naming conventions among MIT students.

3.1.3 Class Name Efficiency Classifications

To further analyze the efficiency of these naming conventions, we defined two major classifications for how students refer to their classes. The first classification, **Name**, refers to instances where a class is referred to by its official name or a shortened/nickname version. Within this classification, we further distinguish between using the complete official name (e.g., "Graduate Machine Learning" for 6.790, "Managerial Finance" for 15.401) and using an abbreviated or colloquial version of the name (e.g., "NLP" for 6.861*, "I am taking Bio" for 7.012). The second classification, **Number**, refers to instances where a class is referred to by its course number. This classification is further divided into using the previous course number (e.g., "6.004", "double-oh-nine") and using the updated course number (e.g., "6.1200", "8.02", "six one twenty"). These classifications serve as the foundation for our analysis of the factors influencing class naming conventions and the efficiency of these conventions in different contexts.

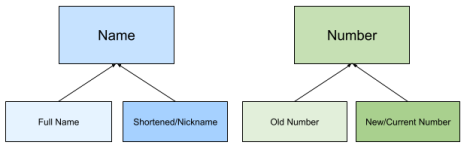

Figure 3: Overview of class naming efficiency classifications used in the study.

3.2 Data Preprocessing and Feature Extraction

Survey responses were preprocessed to ensure data quality and consistency. We removed any incomplete or invalid responses and standardized the formatting of class names and numbers.

For each class mentioned in the survey, we extracted the following features:

- Enrolled number: The number of students enrolled in the class
- Length of name: The character length of the official class name
- Hours: The weekly time commitment for the class
- Rating: The overall rating of the class
- Department: Binary features indicating the department or subject area (e.g., dept_10, dept_11, dept_12, etc.)
- Level: Binary features indicating whether the class is undergraduate (level_U) or graduate (level_G)

These features were selected based on their potential to influence how students refer to their classes and provide insights into the efficiency of class naming conventions. The enrolled number,

length of name, hours, and rating features capture various aspects of the classes that may affect naming preferences. The binary department features (dept_10, dept_11, etc.) allow us to analyze the impact of different subject areas on naming conventions, while the binary level features (level_U and level_G) help us investigate the differences between undergraduate and graduate classes.

By focusing on these specific features, we aim to capture the most relevant information for our analysis while ensuring a manageable scope for the study. These features serve as the foundation for our investigation into the factors influencing class naming conventions and the efficiency of these conventions in different contexts.

3.3 Analytical Methods

To analyze the patterns and trends in class naming practices among MIT students and investigate the factors influencing these conventions, we employed two main analytical methods: logistic regression and neural networks.

3.3.1 Logistic Regression

Logistic regression is a statistical method used for binary classification tasks, making it well-suited for our study, where we aim to predict whether a class is referred to by its name or number based on various features. We trained a logistic regression model using the scikit-learn library in Python, with the following key parameters:

- L2 regularization to prevent overfitting
- Tolerance for stopping criteria set to 0.0001
- Inverse of regularization strength (C) set to 1

The logistic regression model allows us to examine the coefficients associated with each feature, providing insights into the relative importance and direction of influence of different factors on the class naming conventions.

3.3.2 Neural Networks

To further explore the relationships between class features and naming conventions, we implemented a basic neural network model using the PyTorch library. The architecture of our neural network consists of the following components:

- Three fully connected linear layers with (input activation size, output activation size) of (37, 128), (128, 256), and (256, 128)
- ReLU activation function applied after each linear layer
- A final output layer with a single output neuron and sigmoid activation function

The model was trained using binary cross-entropy loss and stochastic gradient descent (SGD) optimizer with a learning rate of 0.01. We trained the model for 10 epochs, using a batch size of 16 and shuffling the training data at each epoch.

The neural network approach allows us to capture potential non-linear relationships between the input features and the target variable (class naming convention). By comparing the performance of the neural network model with the logistic regression model, we can assess the complexity of the underlying patterns in the data and determine which approach is more suitable for this task.

In the following sections, we will present the results obtained from both the logistic regression and neural network models, evaluating their performance using various metrics such as accuracy, precision, recall, and F1-score. We will also discuss the insights gained from analyzing the feature coefficients in the logistic regression model and the implications of these findings for understanding the efficiency of class naming conventions at MIT.

4 Results

In this section, we discuss the results of the binary classification task on MIT's class naming schemes with two main methods described—logistic regression and basic neural network models.

4.1 Result: Logistic Regression

As part of the data analysis, we computed accuracy, precision, recall, and F-1 score generated from our logistic regression model. The summary table of the results is shown in Figure 4.

Model Performance			
Accuracy	**Precision TP / (TP+FP)**	**Recall (Sensitivity) TP / (TP+FN)**	**F-1 Score TP / (TP + 1/2*(FP+FN))**
72.86%	66.67%	84.85%	74.67%

Figure 4: Model performance summary table for the logistic regression model.

Overall, our logistic regression model achieves 72.86% accuracy. Furthermore, relatively high recall (84.85%) compared to precision (66.67%) shows that our model is more likely to make false-positive predictions over false-negative predictions.

Since the labeling scheme of our data is such that 0 is the class being referred to as numbers and 1 is the class being referred to as names, this result shows that our model is more likely to classify the class being referred to as names than the true population. This phenomenon can also be observed in the confusion matrix in Figure 5.

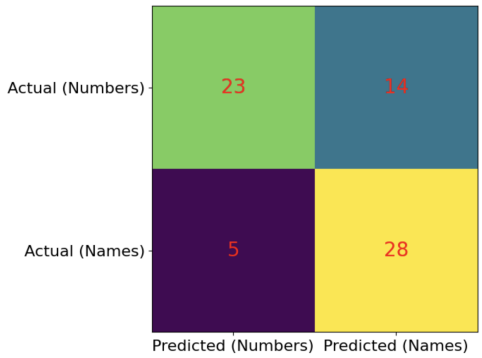

Figure 5: Confusion matrix for the logistic regression model.

4.2 Result: Neural Network

As the second method to analyze our data, we used a basic matching learning model mainly composed of linear and activation layers. For the hyperparameters, the learning rate of 0.005 and epoch size of 10 were used. Validation accuracy and validation loss over the 10 epochs are visualized in Figure 6 and Figure 7, respectively.

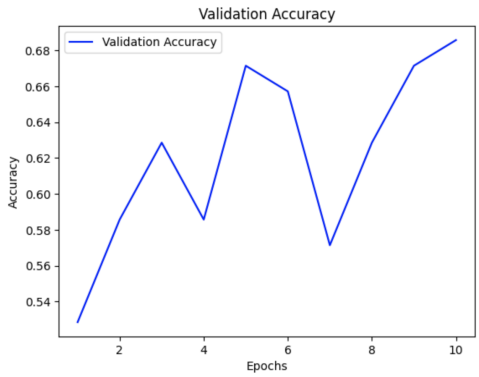

Figure 6: Validation accuracy for the neural network model.

At the end of 10 epochs of training, our neural network model achieved an accuracy of 68.57%. This accuracy, when compared with 72.86% by our logistic regression model, shows that the neural network model may not be the ideal way to analyze our data in this study.

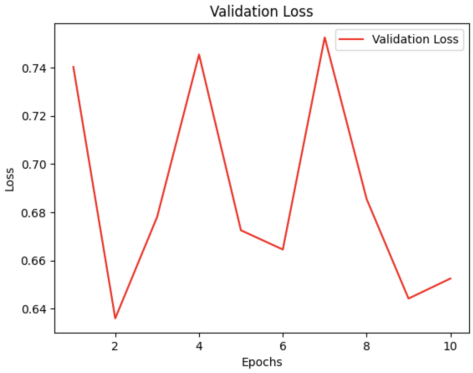

Figure 7: Validation loss for the neural network model.

4.3 Result: Test Set Distribution

It is important to put the 72.86% performance of our logistic regression model in context. Specifically, we want to see if our model outperformed a blind guesser. Here, a blind guesser refers to a prediction system where it predicts the label that is represented most frequently in the dataset. For example, in the testing dataset where 75% of the data are labeled 0 and 25% of the data are labeled 1, a blind guesser could achieve 75% accuracy by simply labeling every datapoint with label 0. To see how our model performed against this hypothetical blind guesser, we look into the test set label distribution, as well as our model's prediction set label distribution. Both of these distributions are shown in Figure 8 and Figure 9.

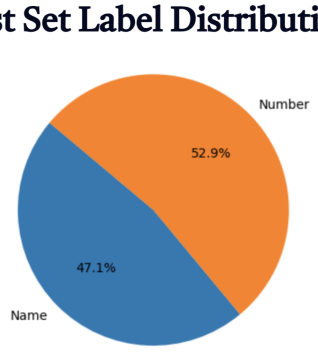

Figure 8: Validation accuracy for the neural network model.

There are two main conclusions we can draw from these two graphs. First, our model demonstrates significant superiority over a hypothetical blind guesser that would have achieved 52.9% accuracy, which is equivalent to the majority label's proportion in the test dataset. Second, although more classes are referred to as numbers than their names in the testing dataset, the opposite is true in

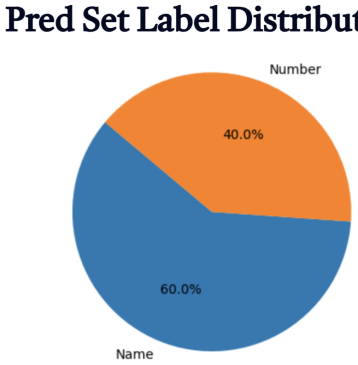

Figure 9: Validation loss for the neural network model.

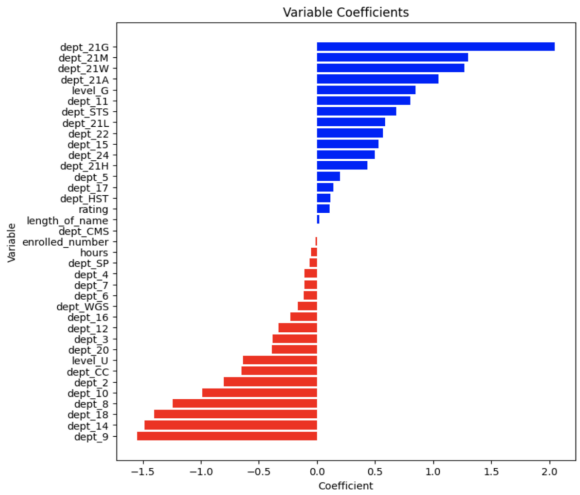

Figure 10: Validation loss for the neural network model. On the y-axis are 37 features associated with each class, and on the x-axis are the coefficients associated with each feature.

the predictions made by our model. This indicates that our model's predictions are driven by learned features from the training process, rather than simply mirroring dataset distribution predictions.

5 Discussion

In this section, we dive deeper into the results produced by our model. Due to its superior performance, we focus our analysis and discussion on the results produced by the logistic regression model. Specifically, we present a feature analysis that shines the light to lingual efficiency present in MIT's classing naming scheme, as well as limitations and future work associated with our study.

5.1 Feature Significance Extraction

With the logistic regression model, we can observe how each feature influences the decisions made by the model by investigating the coefficients associated with each feature (Figure 10).

Since we label the classes that are referred to as numbers with 0s and the classes that are referred to as names with 1s, we can attribute the features with positive coefficients as the factors that indicate name-referred classes and the ones with negative coefficients as the factors that indicate number-referred classes. In the subsections below, we analyze each of these classes and show how they may indicate the presence of language efficiency within MIT's class naming system.

5.1.1 Features for "Name-Referred" Classes

In Figure 11, we can see the features with the 10 most positive coefficients, which is equivalent to the features that contribute the most to the classes being referred to by their names.

Two primary factors contribute to classes being referred to by their names: affiliation with the humanities, arts, and social sciences (HASS) department, and classification as graduate-level courses. As seen in Figure 11, the prominent departments influencing this trend include 21G (Global Languages), 21M (Music and Theater Arts), 21W (Comparative Media Studies), and 21A (Anthropology) at MIT. Since these classes are often not mandatory for graduation, not all students may know their numeric representations. Additionally, the longer prefixes of these department codes (e.g., 21G vs. 5) result in longer class identifiers, increasing complexity in their numbered forms compared to classes from other departments. Graduate-level courses, like those in the HASS department, are typically optional and may be less familiar to students. Therefore, referring to them by their names rather than numbers may help mitigate information loss during communication.

5.1.2 Features for "Number-Referred" Classes

In Figure 12, we can see the features with the 10 most negative coefficients, which is equivalent to the features that contribute the most to the classes being referred to by their numbers.

Two primary factors contribute to classes being referred to by their numbers: affiliation with popular scientific departments and classification as undergraduate-level courses. Notably, departments such as 9 (Brain and Cognitive Science), 14 (Economics), 18 (Mathematics), and 8 (Physics) have a significant influence on this trend, as they

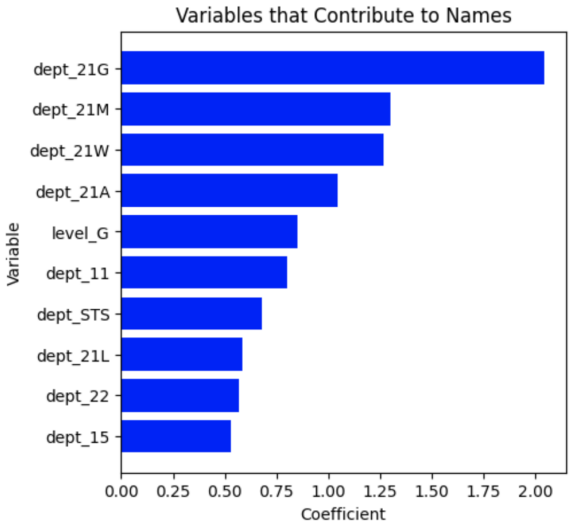

Figure 11: 10 variables that contribute the most to classes being referred to by their names.

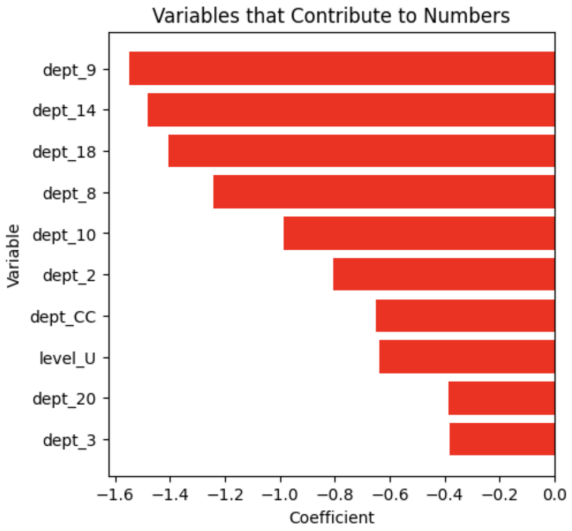

Figure 12: 10 variables that contribute the most to classes being referred to by their numbers.

attract a large number of students. Furthermore, two of these departments feature multiple General Institute Requirement (GIR) classes, mandatory for graduation, making numeric communication more efficient for students. Additionally, undergraduate-level classes tend to have higher enrollment rates and student interest, making numeric representation sufficient for effective communication with minimal information loss.

5.2 Numerical Features Contribution

The features discussed in the previous section were mainly binary (Undergrad/Grad, Department X or not Department X, etc.). In this section, we analyze numerical features. Specifically, we look at 4 features—class rating, length of the official class name, enrollment, and weekly hours commitment—and highlight the key observations made from their associated coefficients as shown in Figure 13.

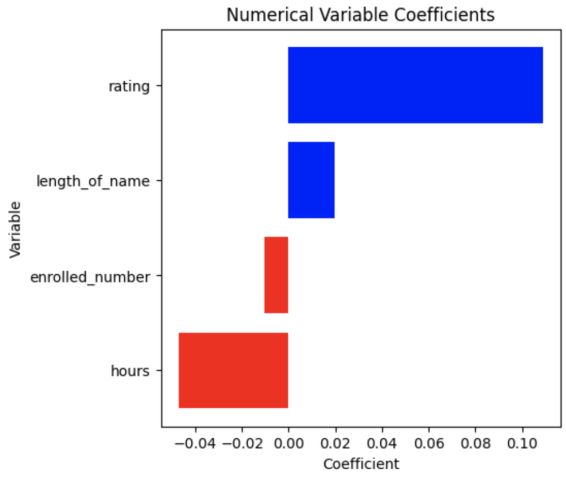

Figure 13: Coefficients associated with 4 numerical features.

Two key observations emerge from this trend, one anticipated and the other unexpected. Firstly, classes with higher enrollment are more likely to be referred to by their numbers. This expected phenomenon arises from the widespread knowledge among students about these classes and their numeric representations, facilitating efficient communication. Conversely, the unexpected trend reveals that longer official class names increase the likelihood of being referred to by their names rather than numbers. Contrary to our hypothesis, longer class names do not lead students to prefer numeric communication. This unexpected behavior suggests that students may opt for shortened or nickname-based references to achieve compression goals, rather than relying on numeric representations.

In addition to these two key observations, it is notable that numerical features exert considerably less influence on the model's decisions compared to binary features. As depicted in Figure 13, the coefficients associated with numerical features typically range between 0 and 0.12. This is in stark contrast to the coefficients for binary features, which range between 0 and 2.1, as illustrated in Figures 11 and 12.

5.3 Model Performance Analysis

As indicated in the Result section, our logistic regression model outperformed our neural network

model. There are two primary reasons for the potential underperformance of our neural network model compared to our logistic regression model. Firstly, prior research suggests that neural network models often require extensive data to surpass traditional data analytic tools (L'heureux et al., 2017). However, our survey was constrained to data from MIT students, encompassing information from only 350 classes, which may not meet the substantial data requirements necessary for the neural network model to realize its full performance potential. Additionally, suboptimal architectural design or hyperparameters may have contributed to the neural network model's underperformance. Despite employing grid search to identify optimal hyperparameters within our search space, there exist unexplored layer designs and hyperparameters that could potentially outperform our model.

5.4 Limitations and Future Work

While our study offers valuable insights into the effectiveness of class naming practices at MIT, several limitations warrant further investigation. Firstly, while our project suggests the potential linguistic optimality of MIT's class naming scheme, formal validation is lacking. Exploring how MIT's class naming maximizes optimization metrics such as Information Bottleneck (IB) as in Zaslavsky et al. (Zaslavsky et al., 2018) could strengthen our findings. Additionally, a deeper dive into feature analysis could yield crucial insights. This entails extracting additional class features (e.g., recitation frequency, website URLs, class age) and exploring correlations between them to enhance model performance. Augmenting data collection efforts through expanded surveys or alternative sources, coupled with employing diverse modeling techniques, may uncover more profound observations.

6 Conclusion

In this study, we investigated the efficiency of language naming conventions for classes at the Massachusetts Institute of Technology (MIT). By employing a mixed-methods approach that combined survey data, computational analysis, and data visualization techniques, we uncovered underlying patterns and relationships between class attributes and naming conventions. Our findings contribute to the broader understanding of language evolution and efficiency in specialized academic domains. By extending the application of information-theoretic principles to the domain of class naming conventions at MIT, we have shown that the drive for efficient communication plays a significant role in shaping naming practices within this unique academic setting. The insights gained from this study have potential applications in optimizing communication strategies, designing effective course cataloging systems, and fostering a shared understanding among students and faculty.

While our study has limitations, such as the need for formal validation of linguistic optimality and the potential for more extensive data collection and feature analysis, it provides a solid foundation for future research in this area. As we continue to explore the factors that shape language evolution and efficiency, we can develop more effective strategies for communication, knowledge sharing, and collaboration within academic institutions and beyond.

Acknowledgments

- We would like to thank Jacob Andreas for all the guidance throughout this project

References

Bevil R Conway, Sivalogeswaran Ratnasingam, Julian Jara-Ettinger, Richard Futrell, and Edward Gibson. 2020-02. Communication efficiency of color naming across languages provides a new framework for the evolution of color terms. *Cognition.*, 195.

Nick C Ellis. 2008. The dynamics of second language emergence: Cycles of language use, language change, and language acquisition. In *The modern language journal*, volume 92, pages 232–249. Wiley Online Library.

Ramon Ferrer i Cancho and Ricard V Solé. 2003. Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences*, 100(3):788–791.

Simon Kirby, Hannah Cornish, and Kenny Smith. 2008. Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *Proceedings of the National Academy of Sciences*, 105(31):10681–10686.

Alexandra L'heureux, Katarina Grolinger, Hany F Elyamany, and Miriam AM Capretz. 2017. Machine learning with big data: Challenges and approaches. *Ieee Access*, 5:7776–7797.

Maciej Pokornowski. 2015. The fourth v, as in evolution: How evolutionary linguistics can contribute to data science. *Theoria et Historia Scientiarum*, 11:45–62.

Noga Zaslavsky, Karee Garvin, Charles Kemp, Naftali Tishby, and Terry Regier. 2022. The evolution of color naming reflects pressure for efficiency: Evidence from the recent past. *Journal of Language Evolution*, 7.

Noga Zaslavsky, Charles Kemp, Terry Regier, and Naftali Tishby. 2018. c. *Proceedings of the National Academy of Sciences*, 115(31):7937–7942.

George Kingsley Zipf. 1949. *Human behavior and the principle of least effort: An introduction to human ecology*. Addison-Wesley Press, Oxford, England.